DOCUMENT RESUME

ED 143 679 TM 006 423

TITLE	District Evaluator's Handbook of Selected Evaluation Procedures. Reprint 1976.
INSTITUTION	New York State Education Dept., Albany. Bureau of School Programs Evaluation.
PUB DATE	76
NOTE	75p.
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	Academic Achievement; Behavioral Objectives; Compensatory Education Programs; Data Analysis; Elementary Secondary Education; *Evaluation Methods; *Evaluators; *Manuals; Measurement Techniques; *Program Evaluation; Program Planning; Sampling; *School Districts; *Statistical Analysis

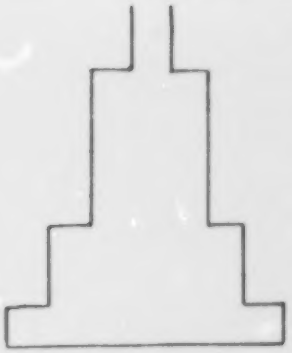
ABSTRACT

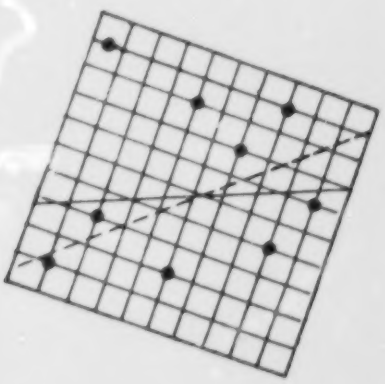
This evaluation handbook was written for school district personnel who may be either novices or experts in the use of educational research techniques. Tailored to categorically aided projects, the handbook describes techniques that will isolate specific activity effects as reflected by pupil achievement. Chapters are devoted to the construction of behavioral objectives, the development and application of a sampling plan, evaluation designs and management of data collection activities, and the application of data analysis techniques. An appendix provides illustrations of evaluation designs applied to Title I (Elementary and Secondary Education Act) projects and a flowchart outlining the steps in planning an evaluation. (EVH)

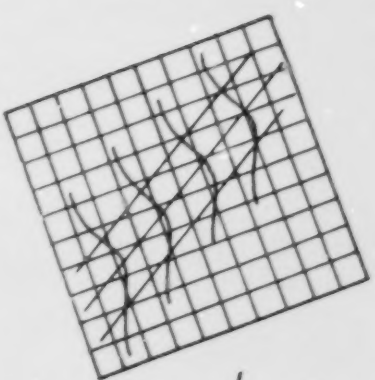
****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

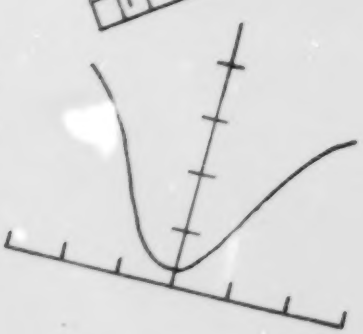
ED143679

TM006 423









U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

District Evaluator's Handbook of Selected Evaluation Procedures

PERMISSION TO REPRODUCE THIS [redacted] MATERIAL HAS BEEN GRANTED BY
The State Education Dept., Albany, N.Y.
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

Reprint
1976

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Bureau of School and Categorical Programs Evaluation
Albany, New York 12234

THE UNIVERSITY OF THE STATE OF NEW YORK

Regents of The University (with years when terms expire)

1981 Theodore M. Black, A.B., Litt.D., LL.D., Pd.D., D.C.L., L.H.D.
Chancellor . . . . . Sands Point
1987 Carl H. Pforzheimer, Jr., A.B., M.B.A., D.C.S., H.H.D.
Vice Chancellor . . . . . Purchase
1978 Alexander J. Allan, Jr., LL.D., Litt.D. . . . . Troy
1981 Joseph C. Indelicato, M.D., L.H.D. . . . . Brooklyn
1986 Kenneth B. Clark, A.B., M.S., Ph.D., LL.D., L.H.D., D.Sc. . . . Hastings on Hudson
1983 Harold E. Newcomb, B.A. . . . . Owego
1988 Willard A. Genrich, LL.B., L.H.D., LL.D. . . . . Buffalo
1982 Emlyn I, Griffith, A.B., J.D. . . . . Rome
1977 Genevieve S. Klein, B.S., M.A. . . . . Bayside
1981 William Jovanovich, A.B., LL.D., Litt.D., L.H.D. . . . . Briarcliff Manor
1983 Mary Alice Kendall, B.S. . . . . Irondequoit
1984 Jorge L. Batista, B.A., J.D. . . . . Bronx
1982 Louis E. Yavner, LL.B. . . . . New York
1979 Laura B. Chodos, B.A., M.A. . . . . Clifton Park
1980 Martin C. Barell, B.A., I.A., LL.B. . . . . Great Neck

President of The University and Commissioner of Education
Ewald B. Nyquist

Executive Deputy Commissioner of Education
Gordon M. Ambach

Deputy Commissioner of Elementary, Secondary, and Continuing Education
Thomas D. Sheldon

Assistant Commissioner for Elementary, Secondary, and Continuing Education Research and Planning
John W. Polley

Director, ESC Education Program Planning
Mark B. Scurrah

Director, Division of Research
Carl E. Wedekind

Chief, Bureau of School and Categorical Programs Evaluation
Leo D. Doherty

FOREWORD

This handbook was written for district personnel who may be either novices or experts in the use of education research techniques. Tailored to categorically aided projects, the handbook is filled with illustrations created for typical projects and contains some techniques that will isolate specific activity effects as reflected by pupil achievement.

The contents of the handbook were assembled in a format that outlines application only. If a coordinator needs to review statistical concepts such as the theory of the characteristics of the normal curve, random events and probability, and parametric and nonparametric tests, he is advised to obtain one of the references included in the bibliography of this handbook. The handbook does not develop concepts underlying inferential statistics.

The handbook provides selected applications as they seem relevant to the construction of behavioral objectives, the development of defensible sampling plans, and the analysis of data collected under definable evaluation designs. In addition, an appendix provides both actual illustrations of evaluation designs currently being applied to Title I projects and an evaluation flow chart for planning.

Local school district personnel charged with the evaluation of compensatory aid programs occasionally require assistance in order to complete the mandated process of evaluation design and data analysis. Such assistance is sometimes secured within the district staff from mathematics teachers, guidance counselors, school psychologists, and other staff who have had training in testing, measurement, and statistics.

External resources which are available for assistance include the BOCES Centers; local universities, especially those in the State University and City University systems; and education research organizations, either university-based or independent. Various university departments, particularly educational or general psychology, guidance, or research and evaluation may be of service in the design of appropriate evaluation procedures. Graduate students registered in such departments have been so employed, but only upon the personal recommendation of recognized, competent faculty members. Experienced coordinators from other districts are sometimes available to provide assistance.

The Department's Bureau of School and Categorical Programs Evaluation can help with the general construction of evaluation designs.

In the present handbook, Appendix D presents an Evaluation Flow Chart for project planning. Certain of the steps indicated (particularly needs assessment) are the definite responsibility of the coordinator. In some of the succeeding steps, the coordinator should be able to provide raw achievement or monitoring data to whoever will be fulfilling the evaluation procedures. The approach is especially important if the school district contracts for an outside evaluation. Contractors are used most efficiently when their efforts are limited to constructing sampling plans, developing evaluation designs, and performing data analysis services, rather than collecting raw data.

Accountability has tremendous importance in the field of compensatory education. Achievement evaluation, a tool of accountability, is being continuously reshaped as processes develop to analyze the effects of categorically aided programs. This handbook is devoted entirely to evaluation procedures tailored to such programs.

The material was compiled and written by George A. Cronk, Jr., associate in education research in the Bureau of School and Categorical Programs Evaluation. During the editing, Robert F. Miller, William Jaffarian, Lee Wolfe, Donald White, and Paul Hayford contributed valuable suggestions and counsel.

TABLE OF CONTENTS

		Page
CHAPTER I:	CONSTRUCTION OF BEHAVIORAL OBJECTIVES	1
CHAPTER II:	DEVELOPMENT AND APPLICATION OF A SAMPLING PLAN . .	3
	Defining the Target Population	3
	Using a Table of Random Numbers	4
	A Quick Method for Approximating the Needed Treatment Group Sample Size When a Nontreatment Group Will Be Used for Comparison	6
	Determining Approximate Sample Sizes Based Upon the Desired Degree of Association for Uncorrelated Samples	8
	Estimation of a Required Sample Size When Testing Treatment Means	10
CHAPTER III:	METHODOLOGY AND MANAGEMENT PLAN	12
	Developing an Evaluation Design	12
	Scheduling and Managing Data Collection	14
	Specifying the Instrumentation	14
CHAPTER IV:	APPLICATION OF DATA ANALYSIS TECHNIQUES	16
	Describing Change Through Descriptive Statistics .	16
	Interpretations of Norm Scores	19
	Standard Scores	19
	Stanine Scores	20
	Percentile Ranks	20
	Large Population Statistical Analysis	21
	$\overline{Z}$ Ratio Applied to Uncorrelated Stanine Means, Posttest Only	21
	Using a t-Ratio for Related Measures on Standard Scores and Percentile Ranks for a Real v. Anticipated Gain Design	24
	Small Sample Statistical Analysis	27
	Applying a t Ratio to the Difference Between a Pretest and Posttest (Correlated Sample) . . .	27
	Actual Posttest Comparison to the Predicted Posttest Scheme of Data Analysis Using a t Ratio .	29
	Applying a t Test to the Difference Between Two Posttests for Independent Samples	32
	Analysis of Variance (ANOVA)	33
	Interpretation of Decision Making	40
	Analysis of Covariance	41
	Summary for Analysis of Covariance (ANCOVA)	47
	The Median Test for Two Correlated Samples	47
	The Median Test for Two Independent Samples	49
	Wilcoxon Matched-Pairs Signed-Rank Test for Two Correlated Samples (N<25)	50
	Chi Square (X^2)	52

Describing Relationships Through Statistical Correlations 54
Pearson Product Moment Correlation Coefficient . 54
Use of the Point Biserial Correlation 56
Use of the t Statistic to Account for the Treatment Impact 59
Comments to Coordinators 61

APPENDIX A - Instructional Activity 62

APPENDIX B - Support Services 64

APPENDIX C - Instructional Activity (Summer) 66

APPENDIX D - Evaluation Flow Chart for Title I Project Planning 67

BIBLIOGRAPHY 68

CHAPTER I: CONSTRUCTION OF BEHAVIORAL OBJECTIVES

There are at least five separate facets of project proposal evaluation plans that must be addressed by project proposal writers. The Bureau of School and Categorical Programs Evaluation reviews the project's (1) objectives, (2) sampling plans, (3) design, (4) data analysis techniques, and (5) plan of presenting the effects of the special learning treatments (activities). If any one of the area is found wanting, a recommendation to disapprove that project is automatically sent to the appropriate approving office.

Objectives. An affirmative answer to the following three questions is prerequisite to the construction of acceptable proposal objectives.[1]

1. Is the objective stated in behavioral terms for the learner? The objective must clearly define what behavioral change (growth) will take place as a result of the treatment.

2. Is the anticipated performance level precisely stated? The proposal writer needs to indicate what degree of change constitutes successful attainment of that objective.

3. Does the objective contain the criteria that define how the reviewer knows that a change has taken place? The means by which evidence of the change will be demonstrated must be included in the objective.

In some cases, the proposal writer may wish to indicate what proportion of the treatment sample will be considered for the successful attainment of the objective.

Following are three illustrations that meet the prerequisites just listed.

AREA OF BEHAVIORAL CHANGE	DEGREE OF CHANGE	CRITERION REFERENCE
Illustration A (Traditional classroom)		
In reading comprehension	the mean of the target population will increase by 1 year	as measured by the Metropolitan Achievement Test.

[1] For a comprehensive approach to framing objectives in behavioral terms, see Preparing Program Objectives: Proposal Guidelines for Categorically Funded Programs, available from The University of the State of New York, The State Education Department, Bureau of School and Categorical Programs Evaluation, Albany, New York, 12234.

Illustration B (Standard evaluation) In mathematical problem solving	the target population will demonstrate achievement beyond expectation[2] ($p \leq .05$)	as measured by the Stanford Achievement Test.
Illustration C (To be used only with criterion refereced treatments)[3] In the mathematical computation of addition	the target population will demonstrate Level 3 mastery	addition of one 5 digit number to another 5 digit number, such as 12345 + 67891, without regrouping.

[2]Expectation as used here means an estimate based upon empirical computation, usually from district regression analysis for the target population, or from prediction based upon individual's regression as described in the real gain v. anticipated gain design discussed later.

[3]At the present time, if a district chose to use Illustration C, the complete set of mastery objectives for every level would have to be submitted with the project application. At some point in the future it is anticipated that the Comprehensive Achievement Monitoring System (CAM) will be refined to the point where a reference by index number will be sufficient specification.

CHAPTER II: DEVELOPMENT AND APPLICATION FOR A SAMPLING PLAN

Defining the Target Population

Although the target population is specified on the application form, additional information is required in the evaluation section of the project proposal.

1. The target population must be defined by the characteristics that will be emphasized in the treatment of the educational deficiency. The prudent district will define as many characteristics of the learners selected for treatment as are feasible. Ultimately, the district will attempt to correlate the particular treatment that is optimal for learners with particular characteristics.

2. Frequently, a compensatory project contains several components. Each component is devoted to different activities for different educational deficiencies. The separate subpopulations by area of treatment must be specified. Districts should also indicate which learners will be included in multiple treatments spanning several components, and which learners will receive only one treatment for one particular educational deficiency. The district is then in a position to determine whether a single effort produces the desired results, or there is a multiplicative effect due to a concerted effort to coordinate several component treatments.

3. When large numbers of pupils (more than 120) are included in a component of a project, analyzing the entire target population for growth as measured by a test is not necessary. Usually, a sample will exhibit the changes taking place in the entire treatment population for that particular treatment. Error in an individual's deviation about a mean will be counterbalanced and the sample will approximate the distribution of a much larger population on a particular characteristic. When any treatment group within a given component is comprised of fewer than 120 pupils, the entire treatment group should be included during the data analysis phase of the project.

4. When a sampling approach is being used by an evaluator, it is critical that the method of sampling be described. When sampling is not done by one of the procedures mentioned below, defensible inferences about the population cannot be made. Verification as to the effectiveness of a treatment is not possible under such circumstances.

Randomized Sampling: The evaluator simply selects students from the treatment population in an aimless or haphazard fashion until he fills the size of the sample sought. The most common course taken in random sampling is to take a table of random numbers and select numbered participants according to the table's "sequence."

Stratified Random Sampling: The evaluator introduces a variable or characteristic to the population and then selects the participants randomly within the subpopulation of that characteristic. For example, consider a New York City reading project where the total treatment group consisted of 1,500 Puerto Rican students and 4,500 Afro-American students. The evaluator desired to obtain a sample of 120 participants. Using the ethnic background as the stratification factor, he would select 30 Puerto Rican students at random and 90 Afro-American students at random. The evaluator obtained a proportional stratified random sample.

Multistage Sampling: The evaluator randomly selects a unit of the population and then samples again within that unit. For example, an evaluator may randomly select several schools from all the schools in his district conducting reading projects with paraprofessionals, and then randomly select grades within each school.

Cluster Sampling: The evaluator purposely clusters schools around one or more factors and then samples within the cluster. Once the clusters are defined the evaluator is free to select randomly or by stratification. For example, an evaluator may want to sample second grade Title I remedial reading students, but by a cost per pupil and size of class situation. He would make a grid or "cell" plan with a cost per pupil axis and a class size axis. After assigning all second grade Title I remedial reading classes to the appropriate "cell" he is free to simply randomly select pupils or to stratify (e.g., by sex, past performance, ethnic origin) his selections if he so chooses.

While sampling plans can be designed to be extremely sophisticated, the basic rule in sampling is this: Keep the sample as free from bias as is possible.

Using a Table of Random Numbers

When an evaluator needs to compare treatment groups to nontreatment (but eligible) groups, the evaluator should select the samples at random. Some evaluators use a roulette wheel, a lottery like the Armed Services draft, or even a basin filled with well-mixed numbered balls or papers. Some evaluators are fortunate enough to be able to assign pupils at random to the treatment group or the regular classroom group at the outset. Other evaluators (with the coordinator) are limited to assigning a treatment to a classroom at random. In each case the main principle that is followed is to select pupils within either the treatment or nontreatment group without regard to an order or system. In other words, every pupil within a treatment group or nontreatment group would have the same chance as every other pupil of being picked to represent the sample. (Actually, as pupils are picked the population shrinks slightly, so that the remaining pupils stand a slightly increased chance [better odds] of being picked).

When selecting random samples, many evaluators use a table of random numbers. Tables of random numbers are usually constructed by computers. Every digit from 0 to 9 that appears in every row or column had an equal chance (with every other digit from 0 to 9) to appear in that spot.

The evaluator can read the numbers consecutively in any direction in the table; that is, horizontally by rows, vertically by columns, or diagonally up or down. The numbers read represent the pupils to be selected for consideration for (1) assignment to a treatment or a regular classroom or (2) selection as test score recipients within a treatment or a regular classroom.

Below is a section of a table of random numbers:

MOCK TABLE OF RANDOM NUMBERS

		COLUMN		
		12345	6-10	11-15
	01	69122	95199	26699
	02	39418	20224	99094
R	03	30033	73090	29531
	04	94068	03488	62386
O	05	06088	39952	26216
	06	60935	83696	06316
W	07	10704	48969	59596
	08	27427	44103	87646
	09	56401	37655	10515
	10	95603	39622	79952

If the target population has less than 100 total pupils eligible for selection, a two digit number is required. (1) Looking at row 01 and columns 1 and 2, the first pupil picked would be pupil #69. (2) Moving horizontally, the second pupil would be pupil #12; the third, pupil #29. If a pupil is picked twice (e.g., row 08, columns 4 and 5 and row 10 columns 10 and 11) simply skip the second entry and move on until the sample is filled. (3) Moving vertically, the second pupil would be pupil #39; the third pupil #30, etc. (4) Moving and dropping diagonally, the second pupil would be pupil #41; the third, #37; the fourth, #34; etc.

If the target population is larger than 100 but less than 1,000 (000 to 999), then three digit numbers are required. If the first pupil was again selected at the starting point of row 01, then columns 1, 2, and 3 are required. Moving horizontally the first pupil is #691; the second pupil is #229; the third pupil is #519.

To find a starting point in the random numbers table a common practice is to take a pencil with the eraser end pointing toward the table, look away from the table, and quickly thrust the eraser onto the table. That number covered by the eraser is the starting point in the table. Another common practice is to roll a pair of dice. Let the digit on one die represent the starting row, and the digit on the other die the starting column. The idea behind the blind thrust or the die throwing approach is to avoid superimposing a "system" of always obtaining the same sequence of numbers.

Illustration: A coordinator was faced with the problem of selecting 20 pupils for a remedial reading treatment from a total population of 70 eligible target pupils. The pupils were on an alphabetical listing. Almost all of the parents of the 70 pupils wanted their children to receive the treatment. The coordinator decided to select the pupils randomly from a table of random numbers. He needed two digit numbers. He would have to disregard any two digit numbers that were larger than 70. Using the blind thrust techniques for starting in the sample table above, the coordinator started at row 04, column 4 and moved horizontally. The pupils were as follows:

68,03,48,~~86~~,23,~~83~~,06,08,~~99~~ ~~73~~,52,
26,21,66,09,35,~~89~~,69,60,~~69~~,16,10,
70,44,~~59~~,~~89~~,59,~~59~~,62

Numbers 86, 83, 99, 73, and 89 are crossed out because they were greater than 70. The second entries for numbers 69 and 59 were crossed out because they had already been removed from the sample.

The coordinator was able to assign the 20 pupils to the treatment class and withstand any charge of favoritism or bias on the grounds of sex, race, creed, etc., from the parents of eligible pupils not receiving the Title I remedial reading treatment.[4]

Illustration: Consider the illustration given above, but assume that a larger target population existed that was composed of several hundred pupils. The 20 pupils receiving treatment were again assigned by use of random numbers from a table. Now, however, a nontreatment pupil scores set up numbering about 20 is required in the spring to compare the treatment group with the nontreatment group for achievement in reading comprehension. The coordinator needs a sample of nontreatment pupil scores since he does not want to use several hundred scores. Option #1: An alphabetical listing of nontreatment pupils is prepared and the same procedure is repeated to yield a random selection of nontreatment pupils for comparison purposes. Option #2: The original list from which the 20 treatment pupils were selected is resurrected, and by continuing on in the table a second set of 20 nontreatment pupils is isolated. (In actual practice, option 2 is usually selected and fulfilled at the same time the random assignment to the treatment group is undertaken.)

A Quick Method for Approximating the Needed Treatment Group Sample Size When A Nontreatment Group Will Be Used for Comparison

The method described below is used to estimate sample size prior to pupil classroom assignment and data collection. The modified McGuigan[5]

---

[4] Since there is an alphabetical bias, this method is far superior to the common practice of systematically selecting every fifth name on an alphabetical class list.

[5] Frank J. McGuigan, Experimental Psychology: A Methodological Approach, 2nd ed., Englewood Cliffs: Prentice-Hall, 1968, p. 364.

approach attempts to answer the question "How many learners in the Title I treatment group and how many comparable learners in the non-treatment group should be tested to be sure to demonstrate significant differences if they exist?" The following steps are suggested:

Step 1. Check last year's Title I treatment group mean and a mean from an equal number of randomly selected eligible nonparticipants. Subtract the nontreatment group's mean ($\bar{X}_2$) from the treatment group's means ($\bar{X}_1$).

Step 2. Calculate the variance for each group separately.

a. If the variances are almost identical, use the following estimation formula:

$$n = \frac{2t^2 S^2}{(\bar{X}_2 - \bar{X}_1)^2}$$

where n = the number of scores in the treatment group

S^2 = the unbiased estimate of the population (common) variance

t = the t ratio for independent means

b. If the variances are considerably different, use the following formula:

$$n = \frac{t^2(S_1^2 + S_2^2)}{(\bar{X}_2 - \bar{X}_1)^2}$$

where S_1^2 = the variance of the treatment group

S_2^2 = the variance of the eligible, nontreatment group

Step 3. Set the probability level for the desired level of significant difference (e.g., $p \leq .05$). On that basis estimate the value of critical t (for $p \leq .05$, be sure to estimate $t > 1.96$; for $p \leq .01$, set $t > 2.58$).

Step 4. Compute the sample size from the formula selected in step 2 above. Illustration: A Title I evaluator was conducting a special computer oriented remedial mathematics treatment for categorically aided learners. Limited funds meant that only 200 pupils out of a target population of 500 pupils were going to be able to receive the special computer oriented treatment. Since parents were extremely sensitive as to whose youngsters would be selected, the pupils were assigned to the treatment or the regular classroom randomly. During the previous year, while there were several implementation problems, the treatment group appeared to have surpassed the regular classroom group in the spring standardized test in mathematics. However, the previous year treatment group was composed of only 50 categorically aided learners. The question before the evaluator was "How many pupils are needed in the treatment groups and nontreatment group to demonstrate a significant difference if there is any?".

First, the evaluator randomly selected a group of 50 eligible nontreatment pupils from the previous years. He arrayed the data as follows:

	Treatment group	Nontreatment group
N	50	50
$\bar{X}$	124	120
S^2	110	125

The evaluator did not know if the two variances (110 and 125) were equivalent (homogeneous). So, he decided to elect formula 2b from above.[6] He arbitrarily selected a t value of 2.1 ($p \leq .05$).

$$N = \frac{t^2(S_1^2 + S_2^2)}{(\bar{X}_1 - \bar{X}_2)^2} = \frac{(2.1)^2\ (110 + 125)}{(124 - 120)^2} = 65$$

In the spring after the districtwide testing, the evaluator will randomly select 70 treatment pupils and 70 nontreatment pupils for mean score comparisons with the t test. The evaluator decided on 70 pupils in each group since the obtained 65 was the very minimum he needed.

Determining Approximate Sample Sizes Based Upon the Desired Degree of Association for Uncorrelated Samples

Hays[7] has described a method for approximating the size of the uncorrelated samples needed when an evaluator wishes to be sure to have enough subjects to make sure significant differences at selected levels of association[8] will show up. The question of determining how many subjects to test in the Title I treatment group and how many eligible, but nontreatment (regular classroom) pupils to include can be answered by this method. This method yields a minimum number.

---

[6] The evaluator could have checked for the homogeneity of the variances by creating an F ratio. That is $F = \frac{\text{larger variance}}{\text{smaller variance}} = \frac{S_1^2}{S_2^2}$. He then would have checked the F table to see if the value there was exceeded by the value resulting from his ratio.

[7] William L. Hays, Statistics for Psychologists, New York: Holt, Rinehart, and Winston, 1963, p. 327.

[8] Sometimes significant differences will show up between measurements taken from some populations, but the level (strength) of the association may be very slight (trivial). The evaluator is interested not only in knowing whether a significant difference existed between groups (and, hence, treatment effects), but also how much of that difference can be associated with given treatments.

1. The evaluator must decide the level of significance that will satisfy his need to reject the null hypothesis (no difference between the treatment and nontreatment group means).

2. The evaluator must decide what degree of association (ω^2 = omega squared) between the treatments and the variance in the obtained scores he desires.

3. The evaluator must solve the following question for delta:

$$\Delta = 2\sqrt{\frac{\omega^2}{1-\omega^2}}$$

ω = the degree of association.

4. The evaluator must solve the following equation for the sample size of the treatment group and then select an equal number for the non-treatment group.

$$n = \frac{2(2.58(p=.01) + 2.33)^2}{\Delta^2} \quad \text{OR} \quad n = \frac{2(1.96(p=.05) + 2.33)^2}{\Delta^2}$$

Illustration: An evaluator desired to test the difference between means obtained from the Spring administration of the districtwide standardized tests in reading for a second grade ESEA Title I class and an equivalent second grade non Title I class. The evaluator desired to assign the pupils at random to the treatment and regular classroom. The evaluator needed to know how many students to test to see if the treatment had the effect that was claimed by the publisher of the materials for the special treatment.

Step 1. The $p \leq .01$ level was selected as the significant difference level.

Step 2. The evaluator desired at least a .30 association between the treatment and the variance in the two groups' achievement scores. (If $\omega^2 = .30$ then the evaluator is inferring that the treatment accounts for approximately 30 percent of the variance in the obtained scores.)

Step 3. $\Delta = 2\sqrt{\frac{.30}{1-.30}} = 1.30.$

Step 4. $n = \frac{2\ (2.58 + 2.33)^2}{(1.30)^2} = 28.5.$

The evaluator needs at least 29 pupils in the Title I treatment group and another 29 pupils in the regular classroom. The evaluator should select a few more[9] pupils in each category than the approximate estimate to be assured of reaching his association if there, indeed, is one of .30. Illustration: Consider the same illustration as above, but let the evaluator decide that he can't obtain 29 pupils in each category because the funds are simply not sufficient to implement the treatment for 29 pupils. The evaluator decides to use his level of significance as .05 instead of .01 as in the previous example, but to retain the association of .30 between treatment and scores. Reapplying Step 4, he has $n = \frac{2(1.96 + 2.33)^2}{(1.3)^2} = 21.78$. Or, for a total target population (treatment plus nontreatment) he needs at least 44 pupils.

In summary, two operations are important when drawing samples and making inferences about categorical aid treatments: (1) the sample must be composed of sufficient numbers to illuminate significant differences when true differences do exist, and (2) given a true difference and corresponding t value, the strength of an association is required for stating that a treatment is important in affecting pupil behavior.

Estimation of a Required Sample Size When Testing Treatment Means

The method described below can be found in greater detail in chapter 12, Sampling and Statistics Handbook for Surveys in Education, prepared and published by the Research Division of the National Education Association, 1965. For the formula to be applied effectively several items of population or sample data are required:

(a) the approximate size of the treatment group (n)

(b) the approximate standard deviation of the group or a previous year's sample (SD) on the variable under study

(c) the approximate error of the mean of the group or of a previous year's sample ($SE_{\bar{X}}$) on the variable under study

In addition, the evaluator needs to select a level of confidence (probability) that will be required at the time of the statistical test. The appropriate deviation value (z) that corresponds to this level is simultaneously determined (e.g., for $p \leq .01$, $Z = 2.58$).

---

[9] A few extra pupils in the samples are advisable since schools receiving categorical aid are noted for attrition in the target population in any year. In rural upstate New York, the exit rate of pupils is estimated at 8 percent while in urban areas the estimate is close to 28 percent.

$$\tilde{n} = \frac{\tilde{SD}^2}{\frac{(\tilde{SE}_{\bar{X}})^2}{z^2} + \frac{\tilde{SD}^2}{N}}$$, where $\tilde{n}$ is the estimated sample size needed.

Illustration: A school district planned to provide 400 (N) remedial fourth grade pupils with ESEA I funded remedial reading treatments. The coordinator wanted to know how many pupils to submit to a pre and post administration of the Metropolitan Achievement Test reading subsections (for a correlated $\bar{z}$ ratio analysis). From an analysis completed the previous year, a similar fourth grade sample (100) of remedial learners has attained a pretest mean of 2.2 (grade equivalent) with a standard deviation (SD) of .4; and standard error of the mean ($SE_{\bar{X}}$) of .04.

The coordinator estimated the random sample size by the formula given above for the proposed statistical test to be interpreted at the .05 level of confidence.

$$n = \frac{(.4)^2}{\frac{(.04)^2}{1.96} + \frac{(.4)^2}{400}} = 134$$

In other words, a sample of 134 randomly selected pupils would represent the district's target population composed of the 400 remedial fourth grade learners. Remember, however, that 134 is the minimum number required and does <u>not</u> allow for pupil mobility in a school year. (Note: Evaluation contractors repeat this estimation procedure for each grade that will be included in an analysis so that inferences in their reports can be made with a stated degree of accuracy and confidence. Coordinators must be prepared to provide the preliminary data so that reasonably close sample sizes can be estimated.)

CHAPTER III: METHODOLOGY AND MANAGEMENT PLAN

Developing an Evaluation Design

A plan of evaluation should be developed before a project is implemented. The purpose of designing an evaluation plan is mainly to be sure that changes (growth) in the learner's behavior can be measured. Measurable behavioral changes provide the educational feedback upon which the improvement of the teaching-learning process depends. Well defined evaluation plans solidify the data collection procedures that finally net data upon which to base defensible decisions. Below are several general evaluation designs appropriate to projects for categorically aided learners.

1. Classic Experimental v. Control. This design is used when two equivalent groups of pupils are going to be compared for a change in behavior. The experimental group receives the treatment while the control group does not.[10]

> Example. A special mathematics computation treatment is to be provided for 30 fourth grade disadvantaged learners who are measured as 2 years below grade level on the New York State PEP Tests. The treatment will be 1 hour per day, 5 days per week for 15 weeks after school in the Title I math lab at the Horace Mann School. A control group, located in the same school with each control student paired with an experimental group student on at least three characteristics will be used to make the comparison. Both groups would be tested on one form of a standardized test before the treatment and then again on another form[11] of the standardized test after the treatment. The results are then compared (see the section on data analysis).

2. Real Gain v. Anticipated Gain (Others). The real v. anticipated gain design is used when a staff can predict the probable

---

[10] Note: In Title I Projects, this design is only permissible when a) funds are so limited that the treatment will not reach all eligible learners or b) when the experimental group receives the special treatment the first half of the year and the control group receives the same treatment during the second half of the year.

---

[11] If a sufficient amount of time has elapsed, and, the nature of the test is such that pupil recall of previous responses is negligible, the same form of the test may be readministered as a posttest.

number of months of achievement for a target population without a specialized treatment. The target population is tested before the treatment and after the treatment and the difference is compared to the anticipated gain.

> Example. Based on their past experience (which was consistent with the Coleman Report), the staff at the John Dewey Elementary School knew, based on last year's class that Miss Lernen's and Mr. Klug's third grade classes would show a reading comprehensive achievement growth of 5 months on the Metropolitan Achievement Test in June. However, this year categorical aid was going to support each classroom with special remedial and developmental reading materials and an education assistant. The target pupils were tested before the treatment and again after the treatment on alternate forms of the Metropolitan Achievement Test. The real gains were compared to the anticipated gains for both classes.

3. Real Gain v. Anticipated Gain (Self). This evaluation design is similar to the preceding design, but depends upon a different prediction for the anticipated gain. The learners in the target population have "averaged" an achievement increment gain to date. The anticipated gain is based upon that increment.

> Example. Mrs. Wissen's class of third grade reading pupils was tested at the beginning of the year. The mean of the scores in vocabulary was converted to show an average monthly gain of .5 months for every month spent in class. Mrs. Wissen anticipated a mean growth of 5 months for a full school year's experience. However, a categorically funded project supplied word attack skill materials, phonics kits, specially tailored enrichment field trip experiences, and an aide. Mrs. Wissen tested the students at the conclusion of the school year and compared the actual monthly average gain to their anticipated monthly gain.

4. Real Gain v. Normalized Gain. This design is appropriate when the evaluator has available a local district norm, State norm, or national norm already established. The target population is tested before and after the treatment as in the previous designs. The difference between the means obtained on the two testings is then compared to the already established norms.

> Example. The Martin Luther King, Jr. School is planning to add a reading laboratory with special materials, remedial reading specialists and educational assistants. Three hundred remedial

learners who scored below two grade "equivalent" levels (below the 23rd percentile on the NYC PEP Test) on reading comprehension are to receive individualized instruction in reading comprehension 1 hour per day, three times per week for 25 weeks. Recently, the Iowa Test of Basic Skills normed two new forms of its tests as part of the nationwide norming process in the district to which Martin Luther King, Jr. School belongs. In other words, a district norm exists.

All target pupils were tested on one form of the Iowa Test of Basic Skills before the treatment and then on an alternate form of the test after the treatment. A simple random sample of 120 target pupils was drawn and the difference between the means obtained by the two testings was compared to the district norm (see the chapter on Data Analysis).

Other designs for evaluating student growth or local variations of the designs above may be applied to a compensatory aid project. Fundamentally, the reasoning behind requiring an evaluation design is to quantify growth exhibited by the learner. With objective data obtained through an evaluation design (a) the learner can receive reinforcement (motivation), (b) a particular treatment can be revised according to empirical findings, and (c) defensible decisions regarding the greatest education "yield" based upon cost and achievement can be implemented with the allocation of future categorical aid.

Scheduling and Managing Data Collection

In addition to the evaluation plan, the schedule of data collection should be specified. The simplest and most widely accepted data collection procedure at the present time is to plan to collect data before the treatment (baseline data) and again after the treatment. When observers are going to collect data during an onsite visit or when questionnaires are going to be released, the time of the year and the time during the project's "life" should be specified. If multiple observations are involved, each observation date should be indicated. Furthermore, each site for each observation should be specified in the project proposal when several schools are included in the same project.

Specifying the Instrumentation

Included in any evaluation design must be some performance to indicate that the behavioral change is exhibited. The most widely accepted means used, presently, is the standardized test. Every project proposal should specify the standardized test that is going to be used for the data collection.

When locally developed instruments are to be used, the instrument or a description of the instrument should be included in the project proposal. Locally developed instruments should be constructed according to accepted procedures for obtaining reliable and valid tests.[12]

Rating scales, observer or pupil checklists, questionnaires, and interview schedules should be constructed so that the responses recorded can be quantified, preferably on an equal interval continuum. This practice becomes critical when correlations between student achievement and selected classroom practices or stimuli are desired. Again, copies or descriptions of the rating scales, checklists, questionnaires, and interview schedules should be attached to the project proposal.

[12]For a succinct practical manual devoted to developing objective tests of achievement that would be appropriate for specialized treatments, see: Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs, N.J.: Prentice Hall, Inc., pp. IX + 118.

CHAPTER IV: APPLICATION OF DATA ANALYSIS TECHNIQUES

The data collected as a result of the project evaluation design will have to be analyzed. The techniques that will be employed in the analysis must be specified in the project proposal. Both descriptive and inferential statistical techniques should be included in the data analysis plan.

Describing Change Through Descriptive Statistics

Such statistics include the mean, median, mode, range, variance, standard deviation, and standard scores. Descriptive statistics are frequently used in compensatory aid projects to indicate where a sample of pupils receiving a special treatment (e.g., remedial reading) would be located relative to all pupils deficient in that educational area.

Definitions:

The mean ($\bar{X}$) is the arithmetic average of the scores obtained by a measurement. The mean is obtained by adding each pupil's score (X_i) to form a population total (ΣX_i) and then dividing by the number of scores (n = pupils).

The median is the point in any distribution of scores where one-half of the scores lie above that point and the other half lie below. A quick approximation to the median can usually be obtained by putting all the scores in consecutive numerical ascending order and counting from the highest score downward until one-half the population is reached.

The range is one plus the difference between the two most extreme scores in the distribution of scores.

The mode is the score that was received most frequently by the target population.

The deviation is the distance on a distribution of scores that indicates how far from the mean a particular score is located.

The population standard deviation is the square root of the sum of every score's deviation from the mean, squared, and divided by the number in the population (n).

$$SD = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}}$$

The variance is the standard deviation squared. If the distribution is normal (bell shaped) then approximately 68 percent of the total population should fall within one standard deviation of the mean.

Ninty-five percent of the total population should fall within two standard deviations of the mean. Ninty-nine percent of the population will fall within three standard deviations of the mean.

A <u>standard score</u> (z_i) is a pupil's deviation divided by the standard deviation $\left(z_i = \frac{X_i - \bar{X}}{SD}\right)$

The mean and standard deviation are the two important parameters (measures) for assessing the central tendency of a distribution. Frequently, remedial learners' scores lie more than one standard deviation below the mean on a standardized test normed (without regard to remedial learners) for a particular grade level. Some districts use this as one criterion for selecting remedial students for a particular treatment funded by categorical aid. Other districts, using locally developed instruments, apply descriptive statistics to establish baseline data for future reference after a treatment has been conducted.

The following example illustrates how to obtain each of the descriptive statistics just defined.

EXAMPLE

On a locally developed word recognition test the following raw scores were obtained from the target population of nine remedial reading pupils($N = 9$): 10, 15, 2, 13, 7, 6, 10, 17, 10

Pupil	Raw Score (X_i)	Deviation ($X_i - \bar{X}$)	Squared Deviation $(X_i - \bar{X})^2$	z Score
X_1	17	+7	49	+1.61
X_2	15	+5	25	+1.15
X_3	13	+3	9	+ .69
X_4	10	0	0	0
median → X_5	10	0	0	0
X_6	10	0	0	0
X_7	7	-3	0	- .69
X_8	6	-4	16	- .92
X_9	2	-8	64	-1.84
	$\Sigma X = 90$	$\Sigma (X_i - \bar{X}) = 0$	$\Sigma (X_i - \bar{X})^2 = 172$	

$\bar{X} = \frac{90}{9} = 10.$

The range was (highest score - lowest score + 1) 16 points.
The mode (most frequently received score) was 10.
The "approximate" median score (midpoint score) was 10.
The mean (arithmetic score) was 10.
Each deviation $(X_i - \bar{X})$ was found by subtracting the mean from the raw score.

Each deviation was squared in the process of finding the standard deviation.[13]

$$SD_{\bar{X}} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N}} = \sqrt{\frac{172}{9}} = \sqrt{19.1} = 4.35$$

The standard score (z score) for each pupil was obtained by dividing each pupil's deviation by the standard deviation.

Theoretically, 68 percent of the pupils should fall within the area of the mean $\pm$ 4.35. This would include students X_3, X_4, X_5, X_6, X_7, and X_8 who all fall in the area from 14.35 down to 5.65 (six out of nine students = 67 percent). The mean ($\bar{X}$ = 10) plus or minus two standard deviations ($\pm$ 2 | 4.35 |) does include all raw scores.

Another useful statistic is the standard error of the mean. This statistic is used for inferential statistical tests. Basically, the standard error of the mean is an estimate of how far the sample mean is from the true mean if the universe of the target population were tested.

$$SE_{\bar{X}} = \frac{SD_{\bar{X}}}{\sqrt{N - 1}}$$

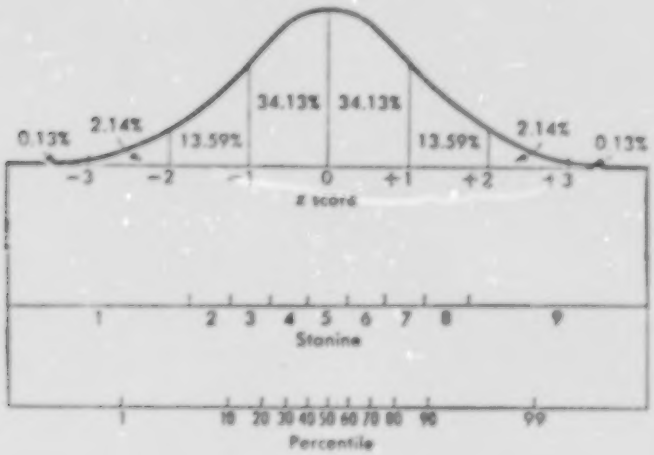
In this example, the standard error of the mean =

$$\frac{SD_{\bar{X}}}{\sqrt{N-1}} = \frac{4.35}{\sqrt{9-1}} = \frac{4.35}{2.83} = 1.53$$

[13] Throughout this chapter, the standard deviation will be "biased." A correction for bias will be introduced when computing the standard error of the mean.

Interpretations of Norm Scores

Below is an illustration of scores most widely reported by standardized achievement tests. The illustration is based upon the distribution of pupils' scores as they relate to the entire population upon which the test was standardized.



Standard Scores

The z score is defined as a standard score. This score for an individual pupil is derived by subtracting the population mean score from the pupil's score and dividing this by the population standard deviation.

$$z = \frac{X_i - \bar{X}}{SD} = \frac{x}{SD}$$

Sixty-eight percent of the normal distribution of scores will lie between a z score of +1.00 to -1.00. ESEA Title I is largely concerned with assisting learners who obtain scores below $z = -1.00$.

z scores can also be used for interpretations on teacher made tests. Illustration: The following sample of scores was obtained from an Afro-American History Test given in five fourth grades.

Pupil	Teacher made test score	$X - \bar{X}$	z
A	3	-3	$\frac{-3}{3.41} = -.875$
B	2	-4	-1.168
C	7	+1	+ .292
D	9	+3	+ .875
E	11	+5	+1.460
F	4	-2	- .584
G	1	-5	-1.460
H	6	0	0.000
I	7	+1	+ .292
J	10	+4	+1.168
($\Sigma n = 10$)	Σ = Sum = 60 $\bar{X}$ = Mean = 6	$\Sigma = 0$	$\Sigma = 0$

$$SD = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} = \sqrt{\frac{106}{9}} = 3.41 \text{ (rounded)}$$

Stanine Scores

Stanine is derived from the contraction of the words standard nine. Standard nine means the normal distribution was divided into nine parts. The mean for the distribution is the midpoint of stanine 5. With the exception of stanines 1 and 9, each band of scores within a stanine is roughly one-half of a standard deviation in width. Below is a chart depicting the percentage of pupils within each stanine and the cumulative number below each stanine.

Stanine	1	2	3	4	5	6	7	8	9
Within (%)	4	7	12	17	20	17	12	7	4
Below (%)	0	4	11	23	40	60	77	89	96

For example, the New York State Pupil Evaluation Program defines being below minimum competence in reading as being below the 4th stanine. This definition encompasses the lowest 23 percent of the normal distribution tail on the left side of the bell shaped curve.

Percentile Ranks

Frequently, standardized tests have a table where raw scores can be converted into <u>percentile</u> points. Percentile points are a value. However, pupils are usually referred to as having fallen at a specific <u>percentile rank</u>, rather than having attained a percentile point. For example, if 78 percent of the norming population attained less than the score value of 20 on a particular measurement device, then the value 20 is the 78th percentile point. A pupil who receives a score of 20 would simultaneously have attained the 78th percentile rank.

One criterion for determining remedial learners in New York State is to survey those pupils who attained a score on the NYC Pupil Evaluation Program Reading Test of the 23rd percentile rank or below.

Large Population Statistical Analysis

When scores from sample populations in excess of 120 are available, one of the easiest methods of statistical analysis is to apply a $\overline{z}$ ratio to the differences between two sets of scores. For a $\overline{z}$ ratio to be significant at the .05 level ($p \leq .05$), a value of $\pm$ 1.96 or greater in magnitude is required. For significance at the .01 level ($p \leq .01$), a $\overline{z}$ value of $\pm$ 2.58 is required.

$\overline{z}$ Ratio Applied to Uncorrelated Stanine Means, Posttest Only

The $\overline{z}$ ratio is defined as $\dfrac{\overline{X}_1 - \overline{X}_2}{SE_{D_M}}$

where $\overline{X}_1$, $\overline{X}_2$ are different samples' means. (Uncorrelated refers to scores or means from two different samples composed of two different sets of individuals. Another formula is used for a pretest-posttest analysis for two sets of scores yielding a pretest mean and/or posttest mean for the same individuals.)

SE_{D_M} is defined as the standard error of the difference between the uncorrelated means. The $SE_{D_M} = SE^2_{M_1} + SE^2_{M_2}$

Illustration:

Two elementary remedial learner schools containing two fourth grade classes were eligible for Title I funded reading activities (the pupils had scored at the 23rd percentile rank or below on the NYS PEP Test the previous year). However, the ESEA Program Office directive (stating that the supplementary expenditure must equal or exceed $350 per child) in reality meant that only one of the classes would get a Title I funded remedial reading treatment. The district evaluator planned to randomly sample within the two schools and administer a pretest in early October and a posttest in late June with the same standardized achievement test. The evaluator planned to compare the rates of growth between the school receiving Title I funded treatments and the school not receiving treatments as well as the stanine positions of the two eligible populations at the end of school year.

Unfortunately for the district evaluator, a lengthy teacher "job action" (which was resolved) and a series of bomb scares forced the evaluator to abolish the pretest-posttest evaluation design. The only scores the evaluator was able to obtain were the Stanford Achievement Test reading scores derived from the districtwide June testing program.

One hundred twenty-two (N_1=122) fourth grade pupils from the target classrooms receiving Title I treatment were distributed in the bottom 4 stanines. One hundred forty-four (N_2=144) fourth grade eligible pupils who did not receive Title I treatment were also distributed in the lower 4 stanines. The district evaluator decided to use a $\bar{z}$ ratio to determine whether a significant difference ($p \leq .05$) existed between the two groups. If a significant difference did exist and favored the treatment group, the evaluator could then infer that Title I funds do assist in bringing about (1) increased achievement and (2) achievement beyond what would have occurred in the regular (nontreatment) classroom.

Below is the way the district evaluator analyzed the data.

Treatment Group		Nontreatment Group	
Stanine	Number of Pupils	Stanine	Number of Pupils
1	17	1	35
2	30	2	37
3	35	3	37
4	40	4	35

Step 1: He summed the scores by treatment and nontreatment.

Treatment Sum (Σ) = (1x17)+(2x30)+(3x35)+(4x40) = 342
Nontreatment Sum (Σ) = 360

Step 2: He found the two means. Treatment $\bar{X}_1 = \frac{360}{122} = 2.8$. Nontreatment Mean $\bar{X}_2 = \frac{360}{144} = 2.5$.

Step 3: He found each mean's standard deviation by (a) taking the deviation of each stanine from the mean, (b) squaring the deviation, (c) multiplying the squared deviation by the number of pupils within that stanine, (d) summing the squared deviations by stanine, (e) and applying the formula for the standard deviations discussed under the descriptive statistics section above.

Treatment

Pupils	Stanine	Deviation ($\bar{X}_1$-stanine value)	Deviation Squared	$n_1 x(\text{deviation})^2$
n_1 = 17	1	-1.8	3.24	55.08
n_2 = 30	2	- .8	.64	19.20
n_3 = 35	3	+ .2	.04	1.40
n_4 = 40	4	+1.2	1.44	57.60
				Σ =133.28

$n_1+n_2+n_3+n_4 = N_1 = 122$

$$SD_{\bar{X}_1} = \sqrt{\frac{133.28}{122}} = \sqrt{1.09} = 1.04$$

Pupils	Stanine	Nontreatment Deviation ($\bar{X}_2$-stanine)	(Deviation)^2	n_ix(Deviation)^2
n_1 = 35	1	-1.5	2.25	78.75
n_2 = 37	2	- .5	.25	9.25
n_3 = 37	3	+ .5	.25	9.25
n_4 = 35	4	+1.5	2.25	78.75
N_2 = 144				Σ = 176.00

$$SD_{\bar{X}_2} = \sqrt{\frac{176.00}{144}} = \sqrt{1.22} = 1.10$$

Step 4: He calculated the standard error for each mean.

$$SE_{\bar{X}_1} = \frac{SD_{\bar{X}_1}}{\sqrt{N_1 - 1}} = \frac{1.04}{\sqrt{121}} = .0945 \qquad SE_{\bar{X}_2} = \frac{SD_{\bar{X}_2}}{\sqrt{N_2 - 1}} = \frac{1.10}{\sqrt{143}} = .0921$$

Step 5: He found the standard error of the difference between the two uncorrelated means.

$$SE_{D_M} = \sqrt{(SE_{\bar{X}_1})^2 + (SE_{\bar{X}_2})^2} = \sqrt{(.0945)^2 + (.0921)^2} = .132$$

Step 6: He applied the $\bar{z}$ ratio.

$$\bar{z} = \frac{\bar{X}_1 - \bar{X}_2}{SE_{D_M}} = \frac{2.8 - 2.5}{.132} = \frac{.3}{.132} = 2.27$$

Since the obtained figure of 2.27 was greater than the figure of 1.96 needed for $p \leq .05$, the evaluator was able to infer[14] that the Title I funded treatments were having an impact upon the reading difficulties of the remedial learners in such a way as to bring about achievement beyond that which would have occurred in the regular classroom (as shown by the nontreatment group).

Using a t-Ratio for Related Measures on Standard Scores and Percentile Ranks for a Real v. Anticipated Gain Design

The interpretation of a gain in group scores (where statistically significant differences have been established) provides a strong indicator of the success of a treatment.

In an "ideal case" a control or comparison group is established which is from the same population as the target population. When all other factors have been held constant, the rate of achievement of the control group is assumed to be the same as that which would have been expected of the treatment group without remediation. The posttest scores of the control group can therefore be used as an estimate of the posttest scores of the treatment group without remediation.

In most cases, a control group is not available. The test publisher's norm group may provide a plausible estimate of non-treatment posttest scores

---

[14] Statistical tests of analysis do not prove anything. Analysis of this nature only permits the evaluator to make inferences against the probability of making a correct choice. The larger the $\overline{z}$ ratio the greater is the probability of making the correct choice. In this illustration, the evaluator had two choices as follows: (#1) the means of the two samples were actually identical and the difference between 2.5 and 2.8 was due solely to sampling variations (chance error), or (#2) the two means were far enough apart to demonstrate a true difference. Choice #1 is called the "null" hypothesis by evaluators. In this case, the evaluator had evidence at a probability level of 95 times in 100 that the true difference existed. On that basis, he rejected choice #1 (the null hypothesis) at the .05 level, thereby accepting choice #2. The evaluator then went beyond the data to account for the difference he computed. Since the sample populations were equivalent and met his assumptions about randomness for a universe of poor readers in the fourth grade, he inferred that the Title I funded treatment caused the difference. The presence or absence of the Title I treatment was defined as the independent variable, while the pupils' reading scores are defined as the dependent variable.

The $\overline{z}$ ratio is computed the same way as the t ratio. The use of the $\overline{z}$ ratio automatically means a large sample (N>120) is involved, while the t ratio usually means a smaller sample. The t ratio is frequently used with Student's distribution for critical values, while the z ratio involves values straight from the normal curve. (cf. Guilford, J.P.)

at substantially less cost and effort. Once an estimated posttest score is determined, an evaluator can perform a t-test to ascertain if the posttest mean that was obtained from the treatment group ($\bar{y}$ observed) is greater than the mean that would have been obtained ($\bar{y}$ expected) had the participants received no remedial treatment.

> Example: A district was planning to initiate a remedial reading treatment for third grade pupils in one school. All of the pupils in that grade who scored lower than the 23rd percentile on the screening device used were going to participate in the Title I treatment. No nontreatment comparison group was available.

The Metropolitan Achievement Tests, Primer, Total Reading was administered before and after the treatment.

Test Results / Pupil #	1	2	3	4	5	6	7	8	9	10
Pretest Standard Score	55	55	53	52	52	49	44	41	41	38
Posttest Standard Score	60	64	58	56	57	53	47	48	51	41

Step 1. Compute:

Mean of Pretest score $\bar{X}$ = 48
Mean of Posttest score $\bar{Y}$ = 53.5

Standard deviation of Pretest Scores SD_x = 6.41

Standard deviation of Posttest Scores SD_y = 6.88

Correlation[15] between Pre and Posttest scores = .94

Step 2. Look up the percentile equivalents of mean pretest and posttest standard scores in the norm tables:

pretest percentile equivalent = 18 (standard score 48)

posttest percentile equivalent = 27 (standard score 54)

The pretest percentile score is used to derive the non-treatment posttest expectation. In the absence of special treatment, it would be expected

15/ The computational formula for these statistics can be found in the texts listed in the bibliography. The standard deviation was described on page 20 and the correlation coefficient is described on pages 54-56.

that a group of pupils would maintain its standing relative to the norm group. Thus, the expected posttest score can be found by looking up the standard score equivalent of the pretest percentile (18%) in the posttest norms tables.

$$\text{Expected mean posttest standard score } (\bar{Y}_{expected}) = 50$$

Step 3. Examine the obtained posttest score in relation to the expected score. If the obtained or observed score is larger than the expected score, there may be some reason to believe that the project was effective. A test of the statistical significance of the difference between expected and obtained posttest scores will determine the probability that a difference of that size was the result of chance factors.

This formula is used:

$$t_{n-1} = \frac{\bar{Y} \text{ observed} - \bar{Y} \text{ expected}}{\sqrt{\frac{(SD_x)^2 \ (SD_y)^2 - 2r_{xy} \ SD_x \ SD_y}{N - 1}}}$$

where: N = number of pupils and all other terms are as defined

$$t_{n-1} = \frac{53.5 - 50}{\sqrt{\frac{(6.41)^2 + (6.88)^2 - 2 \ (.94) \ (6.41) \ (6.88)}{9 - 1}}}$$

$$t_{n-1} = 5.7$$

Step 4. Interpret the obtained t-ratio at $p \leq .05$. The observed value of t (5.7) is greater than the critical value on the table at $p \leq .05$ with 9 degrees of freedom (2.262). This is interpreted by the evaluator as a significant difference between expected and observed posttest means. The inference made is that the test results obtained with treatment are greater than those expected without remedial treatment. In cases where there is a large sample size ($N \geq 120$) the obtained t ratio (at the $p \leq .05$ where a t of $\pm$ 1.96 is significant) is interpreted in this manner:

a) if t is negative and larger than -1.96 (e.g.- 2.1), a significant loss was observed.

b) if t is either negative or positive but less than 1.96, no significant gain can be attributed to the treatment. However, the pupils have not fallen further behind their more advantaged peers.

c) if t is positive and greater than 1.96 a significant gain has been observed (i.e., greater than would have been expected to occur by chance).

Small Sample Statistical Analysis

Statistical tests of inference that are applied to small samples ($N \leq 120$) in compensatory aid projects rest upon several assumptions. One primary assumption involved is that the sample available belongs to a larger population (i.e., of remedial learners). Furthermore, a second assumption is that any descriptive statistic obtained from the sample (i.e., the sample's mean reading score on a test) is an estimate of the population's parameter (the true population mean reading score). Since a sample estimate may be slightly different from the population parameter, evaluators demand that the error of the estimate be accounted for. By way of illustration, if a pretest mean in a Title I prekindergarten were obtained on the Peabody Picture Vocabulary Test in November, the evaluator would want to know whether (1) a posttest mean obtained in May was significantly different; or whether (2) the posttest mean was so close to the pretest mean that the error involved in each testing overlapped to the degree that the posttest mean really was the same as the pretest mean. Inferential statistical procedures attempt to answer this question: How far apart do two parameters (i.e., means) have to be before an evaluator can feel "safe" in declaring that a genuine behavioral change due to treatment intervention has occurred?

In the sections below, two types of inferential statistical tests are described. The first type, called the parametric tests, is based upon the assumption that (1) some characteristics within the population are known and that the sample will possess these characteristics (variables) and (2) the distribution of the characteristics is "normal" in the statistical sense. The t test, analysis of variance, and analysis of covariance are parametric tests described below as they may be applied to compensatory aid projects.

The second type of inferential statistical tests mentioned here are called nonparametric. Nonparametric tests are used when (1) little is known about the population distribution or (2) some characteristics are likely to depart from a normal distribution within the population. Included below are the most frequently used nonparametric tests: (chi square (X^2) and variations of the sign test).

Whenever appropriate, the parametric tests should be used in preference to the nonparametric inferential statistical tests.

Applying a t Ratio to the Difference Between A Pretest and Posttest (Correlated Sample)

Illustration

Consider a remedial reading teacher who desired to conduct special field trip excursions to farms with inner city pupils. Words associated with the agrarian dimension of our society seldom came into use in the everyday language of the target population. Her belief was that the inner city pupils would not recognize or comprehend such words until an association was formed.

The remedial reading teacher tailored a word recognition test to the topics to be generated by the field trips. She gave a pretest to a randomly selected number of pupils before the trips, and then gave a posttest after to the same population. The questions before the teacher were: Could the scores obtained by the pupils have occurred by chance - or, did the field trips change the behavior (word recognition) in the target population? The teacher could see that most of the pupils had improved (some pupils much more than others), but she was uncertain as to how much change was enough to assert that the treatment (field trips) was affecting the pupils' learning. The remedial reading teacher decided to test the difference between the pretest group mean and the posttest group mean with a t ratio to see if the difference was only due to chance (testing errors). The total score possible on the test was 10 points.

Ten pupils (N=10) were administered the pretest and posttest. Below are data arranged from the two testings.

Pupil	Posttest	Pretest	Difference(d)	$(d)^2$
1	5	3	+2	4
2	4	4	0	0
3	7	5	+2	4
4	5	2	+3	9
5	8	3	+5	25
6	4	3	+1	1
7	6	7	-1	1
8	3	3	0	0
9	7	2	+5	25
10	6	3	+3	9
N = 10	Σ 55	Σ 35	Σd = +20	Σd^2 = 78
	Mean 5.5	3.5	$\frac{\Sigma d}{n} = \bar{D}$ = 2.0	

The means are 5.5 and 3.5. (The difference between the means is equal to the mean of the differences (2.0)). The sum of the difference was 20, while the sum of the squares of the difference was 78.

$$t = \frac{\Sigma d}{\sqrt{[N\Sigma d^2 - (\Sigma d)^2]/(N-1)}} \quad \text{or} \quad \frac{\bar{D}}{SE_{\bar{D}}}$$

$$= \frac{+20}{\sqrt{[10\ (78)-(20)^2]/(10-1)}} = \frac{+20}{\sqrt{\frac{380}{9}}} = \frac{+20}{\sqrt{42.2}} = \frac{+20}{6.5} = 3.08$$

For correlated samples (same sample population under two observations) the degrees of freedom = df = N-1 = 9.

The critical value of t for 9 degrees of freedom is 2.262 at $p \leq .05$. Since the obtained 3.08 is greater than 2.262, a significant difference exists between the pretest and posttest scores. The teacher can infer that the difference may have occurred as a result of the treatment. (Without a control group for comparison, the teacher cannot be as certain in this inference.)

Actual Posttest Comparison to the Predicted Posttest
Scheme of Data Analysis Using a t Ratio

Real (treatment posttest) v. anticipated (without treatment) posttest design.

Step 1. Obtain each pupil's pretest grade equivalent.

Step 2. Subtract 1 (since most standardized tests start at 1.0).

Step 3. Divide the figure obtained in step 2 by the number of months the pupils have been in school to obtain a hypothetical (historical regression) rate of growth per month. (Ignore kindergarten months. 1 school year = 10 months.)

Step 4. Multiply the number of months of Title I treatment by the historical rate of growth.

Step 5. Add the figure obtained in step 4 to the pupil's pretest grade equivalent (step 1).

Step 6. Test the difference for significance between the group predicted posttest mean and the obtained posttest mean with a correlated t test.

In September, a diagnostic reading teacher administered the Metropolitan Achievement Test (as a pretest) to 30 fourth grade learners who had scored below minimum competency on the New York State Reading PEP Test.

The 30 pupils participated for the first time in an ESEA Title I remedial project conducted from the first week in October through the last week in May (treatment time = 8 months). The reading diagnostician re-administered an equivalent level form of the Metropolitan Achievement Test (as a posttest) during the first week of June to the 30 pupils.

From the September (pretest) administration, the diagnostician calculated the individualized predicted June scores based upon the pupils' historical rate of gain (using the method described in steps 1 through 4 above) that would have been anticipated if the ESEA Title I treatment had not intervened in addition to the regular classroom reading instruction. The diagnostician then compared the predicted posttest scores to the actual posttest scores by the statistic called the t test (critical ratio) to determine whether the 30 pupils' achievement was beyond expectation.

Pupil	Pretest	Posttest Predicted	Posttest Actual	Difference	Difference Squared
1	2.5	2.9	3.2	+ .3	.09
2	2.8	3.3	3.5	+ .2	.04
3	2.2	2.5	2.6	+ .1	.01
4	1.8	2.0	2.0	0	.00
5	2.9	3.4	3.8	+ .4	.16
6	3.0	3.5	3.9	+ .4	.16
7	2.8	3.3	3.2	- .1	.01
8	2.5	2.9	3.2	+ .3	.09
9	2.3	2.7	2.8	+ .1	.01
10	2.0	2.3	2.8	+ .5	.25
11	2.1	2.4	3.0	+ .6	.36
12	2.7	3.1	3.2	+ .1	.01
13	2.0	2.3	2.5	+ .2	.04
14	2.5	2.9	3.5	+ .6	.36
15	2.4	2.8	2.7	- .1	.01
16	2.2	2.5	2.7	+ .2	.04
17	2.6	3.0	3.2	+ .2	.04
18	2.3	2.7	2.9	+ .2	.04
19	2.2	2.5	3.0	+ .5	.25
20	2.5	2.9	3.7	+ .8	.64
21	2.3	2.7	2.9	+ .2	.04
22	2.8	3.3	3.9	+ .6	.36
23	1.5	1.6	1.8	+ .2	.04
24	2.7	3.1	3.4	+ .3	.09
25	2.3	2.7	3.1	+ .4	.16
26	2.5	2.9	3.2	+ .3	.09
27	2.1	2.4	2.8	+ .4	.16
28	2.2	2.5	3.0	+ .5	.25
29	2.3	2.7	3.6	+ .9	.81
30	2.7	3.1	3.0	- .1	.01
N = 30	SUM	82.9	92.1	+9.2	4.62
	MEAN	2.76	3.07		

The pupils have had 30 months of regular school at the time of the pretest.

Step 1. Pupil #1's pretest score was 2.5.

Step 2. Subtract 1 from 2.5 = 1.5 .

Step 3. Divide 1.5 by 30 (months).

Multiply .05 times the number of months of Title I treatment
.05 x 8 = .4 .

Step 4. Add .4 to (the pretest) 2.5 = 2.9.

This figure is the anticipated posttest score (2.9) for pupil #1.

Repeat for each pupil.

Record each pupil's May posttest score.

Subtract each predicted posttest score from the actual (May) posttest score.

Sum the differences. (Σd)

Square the differences individually.

Sum the squared differences. (Σd^2)

$$t = \frac{\Sigma d}{\sqrt{\left[N \cdot \Sigma d^2 - (\Sigma d)^2\right] / (N-1)}}$$

$$t = \frac{9.2}{\sqrt{\left[30 \cdot (4.62) - (9.2)^2\right] / 30-1}} = \frac{9.2}{\sqrt{\frac{53.96}{29}}} = \frac{9.2}{\sqrt{1.86}} = \frac{9.2}{1.36} = 6.76$$

The degrees of freedom (df) = N-1. Look in the t table under df = 29 for the value of t under columns .05 and .01 (two tailed tests). Since our t of 6.76 is greater than the table value of 2.756, at the .01 level of probability, we may infer that this target population achieved beyond expectation in the Title I funded treatment. In other words, an inference in this illustration is that the pupils did exceed (in reading achievement as measured by this standardized test) what would have occurred in the regular classroom without the special Title I treatment. However, in another illustration, the pupils might all have exceeded their predicted posttest scores, but if the obtained t value was less than the critical value (in this case 4.92, $p \leq .01$), a judgment of no significant difference would have been appropriate. This information might provoke a recommendation for a change in treatment.

Applying a t Test to the Difference Between Two Posttests for Independent Samples

Illustration:

A remedial target population was located entirely in one elementary school. However, since Title I treatments were to be applied to supplement regular offerings at a rate of $350 per pupil, not all pupils within that particular school would be able to receive treatment. The target pupils were randomly assigned to each classroom within grade levels so that no favoritism could be charged to the selection of the pupils for supplementary treatments.

The Title I remedial reading teacher desired to know whether a new curriculum, tailored to the needs of the target population, produced changes in student achievement beyond what would have occurred in the regular classroom. The new curriculum (requiring considerable alteration in teaching method) was being resisted by many of the teachers in that school even though the approach had proven itself experimentally in other schools with similar target populations.

Twenty students (N_1=20) who received the special curriculum in the third grade took a reading subtest from a standardized achievement test in June. The mean was 5.6. The standard deviation was .8. The standard error of the mean was .18.

Twenty students (N_2=20) in another third grade who did not receive the special curriculum took the same reading subtest (identical form and level) in June. The mean for this group was 4.3. The standard deviation was 1.2 and the standard error of the mean was .28.

While the treatment group mean looked larger than the nontreatment group mean, the remedial reading teacher was uncertain as to how much difference in the two means was due to real differences and how much difference was due to chance (testing error). The teacher decided to test the difference between the two means to see if statistical significance did prevail.

Treatment group (N_1=20)	Nontreatment group (N_2=20)
$\bar{X}_1$ = 5.6	$\bar{X}_2$ = 4.3
$SD_{\bar{X}_1}$ = .8	$SD_{\bar{X}_2}$ = 1.2
$SE_{\bar{X}_1}$ = .18	$SE_{\bar{X}_2}$ = .28

The degrees of freedom (df) from noncorrelated (independent) samples = $N_1 + N_2 - 2$. In this case, df = 20 + 20 - 2 = 38.

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{(SE_{\bar{X}_1})^2 + (SE_{\bar{X}_2})^2}} = \frac{4.3 - 5.6}{\sqrt{(.28)^2 + (.18)^2}} = \frac{-1.3}{.333} = -3.91$$

Disregard the minus sign.

The remedial reading teacher wanted to know whether the value of 3.91 could occur by chance 5 times in 100 times (probability = .05, often written as $t_{.05}$).

The teacher turned to the t table in the rear of her statistics book. Looking down the left column for the degrees of freedom (df = 38) and at the top column for the two-tailed test level of significance of .05, the teacher found the critical value of t at the .05 level to be 2.025. Since $3.91 \geq 2.025$, the teacher was fairly confident that there was a significant difference between the two means. The teacher then inferred that the special[16] curriculum contributed to greater achievement for that target population as measured by that standardized subtest than would have occurred in the regular classroom that year.

Analysis of Variance (ANOVA)

Categorical aid coordinators frequently desire to isolate the treatments that brought about the most change in the mean scores of the target pupil samples. When just two treatments are involved, and the target population is assigned at random to the treatments, the t ratio described above is appropriate. However, if several classrooms are using different treatments, a series of t ratios with each treatment mean taken against every other treatment mean two-at-a-time would be extremely laborious and quite possibly isolate too many differences due strictly to chance. By using the analysis of variance (ANOVA), the evaluator has

[16] It is important to remember that outcomes of a treatment may be influenced by the Hawthorne Effect. Similarly, other intervening variables such as teacher skill, time of day of treatment, duration of treatment, repetition of events within treatment, etc. may all have influenced the significant difference obtained in this case. Causes such as these are one of the main reasons why repetitive evaluations have to be instituted in spite of the fact that certain treatments appear to have "proved" themselves effective and valid.

The advantage of (1) simultaneously testing all treatments for significant differences (thereby saving labor), and (2) including all the data within every treatment to make a much closer estimate of the population variance from which the samples were drawn. In other words, the evaluator is applying the null hypothesis (no real differences -- all differences due strictly to sampling) simultaneously to all treatment samples.

ANOVA is a statistical procedure for partitioning the total variance of measures from treatments into components of variance. Measures of samples (ie., means, variance, standard deviations, etc.) are affected by "error" arising from controlled or uncontrolled sources. The evaluator attempts to discover which components of variance arising from uncontrolled (sampling) and controlled effects (treatment) can be accounted for by comparing such variance components with the "variance error." For this purpose an F (Fischer) ratio is formed. The denominator of the F ratio is composed of an error estimate that arises from random sampling and all other sources that are unaccounted for (called residuals). The numerator of the ratio is the estimate of the variance components that arises from the categorical aid treatments.

$$F = \frac{\text{variance from treatment effects}}{\text{error variance from sampling and residuals}}$$

Assumptions concerning the use of ANOVA can be reviewed in Guilford, J.P. Fundamental Statistics in Psychology and Education, N.Y.: McGraw Hill Book Co., 1965, p. 274.

Between - sample sum of squares ---- between sample variance

Each of the treatments has a sample mean ($\bar{X}_s$) that was obtained from the pupils' individual scores (X) within the sample. Each of the sample means deviates from the population mean. ($\bar{X}_s - \bar{X}_t$).

This estimate of the variance will be computed by summing the squares of the deviations of all pupils by sample from the population mean ($\bar{X}_t$), called the "between - sample sum of squares" (SS_b). And then dividing by the number of treatments minus one. This estimate is noted as the between-samples mean square $(MS)_b = \frac{(SS)_b}{k - 1}$

Where k = treatments. The degrees of freedom (df_b) associated with this term is the denominator, or k-1. (Note that the deviation is a sample's mean distance squared from the grand mean).

Within - sample sum of squares ---- within - sample variance

One assumption is that the variances of each of the samples are equal (except for the effects of randomized sampling.) Therefore, the sum of squares of samples should yield an estimate of the population variance.

Each deviation of each pupil from his own sample mean is squared and summed (SS_w). The sum of the squared deviations is divided by the total number of pupils from all samples minus the number of sample treatments. $(MS)_w = \frac{SS_w}{N - k}$. The degrees of freedom (df_w) associated with this term is the denominator, or N-k.

$$F = \frac{(MS)b}{(MS)w} = \frac{\Sigma \frac{(\Sigma X)^2_s}{n_s} - \frac{(\Sigma X)^2}{N}}{\Sigma(\Sigma X)^2_s - \Sigma(\Sigma X)^2_s / n_s}$$

ANOVA - Example

Four classrooms of randomly assigned remedial learners were taught under four different classroom organizations for teaching remedial mathematics.

treatment 1: An aide and a teacher divided the class into small groups and both adults taught the groups separately.

treatment 2: An aide and a teacher selected pupils one at a time for one-to-one tutoring.

treatment 3: An aide did all the noninstructional tasks in the classroom, thereby freeing the teacher for additional small group instruction.

treatment 4: An aide did all the noninstructional tasks in the classroom, thereby freeing the teacher for additional one-to-one tutoring.

The treatment is considered to be the independent variable, while the amount of gain on a standardized test will be considered to be the dependent variable for each pupil.

Step 1: Table the data and compute (a) the sums and means by treatment, (b) the grand sum, and (c) the grand mean.

Treatment
(Gain is in months)

I		II		III		IV	
Subject	Gain(X)	Subject	Gain	Subject	Gain	Subject	Gain
1	9	14	7	26	16	36	16
2	3	15	10	27	8	37	14
3	11	16	7	28	19	38	19
4	17	17	15	29	14	39	17
5	10	18	8	30	12	40	16
6	8	19	6	31	7	41	13
7	8	20	4	32	10	42	12
8	9	21	5	33	15	43	18
9	7	22	3	34	19	44	16
10	15	23	5	35	14	45	14
11	9	24	12			46	23
12	5	25	6				
13	8						
$n_I = 13$	$\Sigma = 119$	$n_{II} = 12$	$\Sigma = 88$	$n_{III} = 10$	$\Sigma = 134$	$n_{IV} = 11$	$\Sigma = 178$

$$\Sigma\, n_I + n_{II} + n_{III} + n_{IV} = N = 13 + 12 + 10 + 11 = 46$$

$$\Sigma\, X_t = \Sigma X_I + \Sigma X_{II} + \Sigma X_{III} + \Sigma X_{IV} = 119 + 88 + 134 + 178 = 519$$

$$\bar{X}_I = 9.15 \qquad \bar{X}_{II} = 7.33 \qquad \bar{X}_{III} = 13.40 \qquad \bar{X}_{IV} = 16.27$$

$$\bar{X}_t = 519/46 = 11.28$$

Step 2: (Instead of using the deviation scores with much opportunity for mathematical error with decimal place manipulation, the raw score will be manipulated to generate the same between sample and within sample variance statistics.)

(a) Multiply $\bar{X}_t$ by $\Sigma X_t = \dfrac{(\Sigma X_t)^2}{N} = 5854.32.$

(b) Square each pupil's score in his respective treatment, and sum for a treatment sum.

$$\Sigma (X^2)_I = 1253 \qquad \Sigma (X^2)_{II} = 778 \qquad \Sigma (X^2)_{III} = 1962 \qquad \Sigma (X^3)_{IV} = 2976$$

(c) Sum the treatment sums obtained in 2(b) above,

$$(\Sigma (X^2)_I + \Sigma (X^2)_{II} + \Sigma (X^2)_{III} + \Sigma (X^2)_{IV}) =$$

$$\Sigma\ 1253 + 778 + 1952 + 2976 = 6959$$

(d) Square the sum of the pupil scores by treatment and divide by the number of pupils in the treatment.

$$(\Sigma X_I)^2/n_I = (119)^2/13 = 1089.31 \qquad (\Sigma X_{II})^2/n_{II} = (88)^2/n_{II} = 645.33$$

$$(\Sigma X_{III})^2/n_{III} = (134)^2/10 = 1795.50 \qquad (\Sigma X_{IV})^2/n_{IV} = (178)^2/11 = 2880.36$$

(e) Sum the treatment sums obtained in 2(d) above.

$$\Sigma (1089.31 + 645.33 + 1795.60 + 2880.36) = 6410.60$$

Step 3: The computations will be entered in the following chart:

Sum of Squares				
Between	(item 2e)	-	(item 2a)	= ?
Within	(item 2c)	-	(item 2e)	= ?

Sum of Squares				
Between	6410.60	-	5854.32	= 556.28
Within	6959.00	-	6410.60	= 548.40

Step 4: Table the remainders computed in step 3 and divide each by its respective degrees of freedom to obtain the variance estimate.

Analysis of Variance

Source of Variation	Sum of Squares		df		Variance Estimate
Between	556.28	÷	3	=	185.43
Within	548.60	÷	42	=	13.01

$$F = \frac{MS_b}{MS_w} = \frac{185.43}{13.01} = 14.25$$

Upon referring to a table with F ratios, locate the critical value [the numerator has 3 df (horizontal line) and the denominator has 42 df (vertical line)] which is 4.64 at the .01 level. Since 14.25 is greater than 4.64, the coordinator can conclude that the treatment (method) affects the pupil's amount of gain in mathematics.

Scheffé Method for Isolating Significant Differences

Once an evaluator has determined that the method of teaching mathematics does make a difference, he is most interested in determining which comparisons yield significant differences. The four steps in the Scheffé method for this purpose involve:

(1) computing F ratios between treatment samples taken two at a time, ie., $F = \frac{(\bar{X}_I - \bar{X}_{II})^2}{MS_w (n_I + n_{II}) / n_I n_{II}}$

(2) locating the critical values of F at .01 or .05 when

$df_1 = K - 1$ and $df_2 = N - K$

(3) calculating F^1. $F^1 = (k-1)$ F (critical value)

(4) checking to see whether obtained $F > F^1$

Step 1: $$F = \frac{(\bar{X}_I - \bar{X}_{II})^2}{MS_w(n_I+n_{II})/n_I n_{II}} = \frac{(9.15 - 7.33)^2}{13.01(13 + 12)/(13)(12)}$$

$$= 3.3124/2.08 = 1.59$$

$$df_1 = K - 1 = 3,\ df_2 = 42$$

Treatment Comparison	I,II	I,III	I,IV	II,III	II,IV	III,IV
F	1.59	7.85	23.29	15.43	35.38	3.33

Step 2: The critical values of F located in the Fischer Table are 2.83 ($p \leq .05$) and 4.29 ($p \leq .01$) when $df_1 = 3$, and $df_2 = 42$.

Step 3: $F^1 = (K-1) \times F$ (critical) where K = the number of treatments.

$$F^1_{(.05)} = 3 \times 2.83_{(.05)} = 8.49$$

$$F^1_{(.01)} = 3 \times 4.29_{(.01)} = 12.87$$

Step 4: Compare the values of the obtained F with F^1.

Treatment Comparison	Obtained F	Significance .05	.01
I,II	1.59	no	no
I,III	7.85	no	no
I,IV	23.29	yes	yes
II,III	15.43	yes	yes
II,IV	35.38	yes	yes
III,IV	3.33	no	no

Interpretation for Decision Making

Statistical. Treatment I appears to be as good as treatment II in bringing about change in pupil behavior in mathematics - but treatment I is apparently inferior in bringing about change when compared to treatment IV. Treatment II is decidedly inferior to treatments III and IV. Treatment III is no better or worse than treatment IV, failed to achieve significance over treatment I, but did achieve a significant difference over treatment II.

Educational. When using teachers to teach with aides serving in a support capacity (noninstructional), the most significant differences appeared (see treatment IV comparisons especially). If the aides are used in a non-instructional fashion while the teachers teach, there appears to be no advantage in small group instruction over one-to-one tutoring. Therefore, the supervisor of mathematics can assign those teachers by teacher preference to teach using either method (small group or one-to-one) without paying a price in mathematics achievement gains.

Evaluation. Judgement over the comparison of treatment I and treatment II might well be suspended. The evaluator might wish to replicate the evaluation study after dropping treatment II entirely for the next year. The comparison of treatments I and III came close (7.85) but failed to achieve significance. For the evaluator's own curiosity, he might wish to apply a combined mean Scheffé comparison to see if the use of aides in certain capacities is decisive.

Combine treatment means I and II; and III with IV.

$$\bar{X}_{I+II} = (n_I \bar{X}_I + n_{II} \bar{X}_2)/(n_I + n_{II})$$

$$= (119 + 88)/25 = 207/25 = 8.28$$

$$\bar{X}_{III+IV} = (n_{III} \bar{X}_{III} + n_{IV} \bar{X}_{IV})/(n_{III} + n_{IV})$$

$$= (134 + 178)/21 = 14.86$$

Compute the F ratio for combined means.

$$F = \frac{(\bar{X}_{I+II} - \bar{X}_{III+IV})^2}{MS_w/(n_I + n_{II}) + MS_w/(n_{III} + n_{IV})}$$

$$= \frac{(8.28 - 14.86)^2}{13.01/(25) + 13.01/(21)} = \frac{(6.58)^2}{.52 + .62} = \frac{43.30}{1.14}$$

$$= 37.9$$

Since the F of 37.9 far exceeds the F^1 of 12.87, the evaluator is reasonably sure that the advice to the district to keep the teachers (alone) teaching and to keep the aides from direct mathematics instruction is to be preferred for maximum pupil gains.

Analysis of Covariance

Evaluators want to insure that differences between treatments are genuinely within the limits of error surrounding the treatment (independent) variables. Sometimes the pupils as a group in a given treatment bias the results because of an uncontrolled causal circumstance [i.e., in New York City some homogeneous grouping of remedial learners yields classrooms of low "exponent" (a euphemism for high I.Q.) pupils]. The situation arises frequently when Title I treatments are applied to entire classrooms which contain whole classes with different starting levels of achievement. When the samples of remedial learners cannot be controlled through random assignment, matching by pairs, etc. a statistical "control" is introduced to "adjust" the two populations so that they can be compared for growth. In other words, the initial level before a treatment for a class may be different, so that an adjustment would have to be made to offset differences in achievement that are attributable to the differences at the initial level. The analysis of covariance is used to remove the bias that favors one class over another at the outset of a treatment.

Basically, the analysis of covariance uses the same principles as the analysis of variance -- but with the addition of products leading to the adjustment of scores.

Illustration: Three target schools containing categorically aided learners are going to receive Title I funds for remedial reading treatments. (No random assignment to treatment was possible for the target population.) Each school will employ a different treatment. Treatment I is the Durrel-Murphy Approach. Treatment II is the Sullivan Approach. Treatment III is the Gattegno Approach. The pupils in each treatment were given the same reading pretest (X_i) and posttest (Y_i). Treatment I contained 14 pupils. Treatment II contained 12 pupils ($N_{II} = 12$). Treatment III contained 10 pupils ($N_{III} = 10$).

Below, the computation table is laid out as was the case in the analysis of variance with the addition of the cross product table. Since there are so many Xs, the summed score totals (previously called X_{totals}) are denoted as T in this illustration.

Treatment

Pupil	I		II		III		
	Pre(X)	Post(Y)	Pre(X)	Post(Y)	Pre(X)	Post(Y)	$N=14+12+10=36$
1	5	8	10	13	6	10	$T_x=140+236+122=498=X$
2	9	11	18	21	8	12	$\bar{X}_x=\frac{498}{36}=13.83$
3	11	13	22	26	13	17	$\frac{T_x^2}{n}=\frac{(498)^2}{36}=6889.00$
4	4	8	8	12	7	11	$\Sigma(\Sigma X^2)=1682+5704+1938=9324$
5	17	19	34	36	25	29	
6	20	23	40	43	23	27	$\frac{\Sigma T_{x_i}}{n}=1400.00+4641.33+1488.40=7529.73$
7	4	7	8	13	15	19	$Ty=180+280+162=622$
8	6	10	14	19	12	16	$\frac{T_y^2}{n}=\frac{(622)^2}{36}=10746.78$
9	8	12	16	18	9	13	$\Sigma(\Sigma Y_k^2)=2560+7520+3074=13154$
10	12	13	24	29	4	8	$\frac{\Sigma T_{y_i}^2}{n}=2314.29+6533.33+2624.4=11472.02$
11	11	14	22	25			$T_{xy}=2056+6524+2426=11,006$
12	10	12	20	25			$\frac{\Sigma(T_{x_i})(T_{y_i})}{n1}=\frac{(140)(180)}{14}$
13	13	15					$+\frac{(236)(280)}{12}+\frac{(122)(162)}{10}$
14	10	15					9283.17
n_i	14		12		10		$\frac{T_xT_y}{n}=\frac{(498)(622)}{36}=$
T_{x_i}, T_{y_i}	140	180	236	280	122	162	8604.3
$\bar{X}$ pre, $\bar{Y}$ post	10.00	12.86	19.67	23.33	12.20	16.20	$Tx_y-\frac{Tx\ Ty}{n}=2401.7$
$\Sigma X^2\ \Sigma Y^2$	1682	2560	5704	7520	1938	3074	
$\Sigma(X_i)(Y_i)$ $=T_{x_iy_i}$	2056		6524		2426		
$\frac{T_{xi}^2}{n}\ \frac{T_{yi}^2}{n}$	1400.00	2314.29	4641.33	6533.33	1488.40	2624.40	

Step 1 - Array the data as in the chart.[17]

Step 2 - Sum pretest and posttest scores in columns.

	X	Y	X	Y	X	Y
Tx_i, Ty_i	140	180	236	280	122	162

Step 3 - Square every X score individually by treatment and sum across all three treatments.

$1682 + 5704 + 1938 = 9324$

Step 4 - Add the sums of the X columns ($\overset{\text{grand}}{\text{sum}} = T_x$).

$140 + 236 + 122 = 498 = T_x$

Step 5 - Square the grand sum of Xs and divide by the total population (N).

$$\frac{(498)^2}{36} = 6889$$

Step 6 - Subtract the result of step 5 from the sum obtained in step 3.

$9324 - 6889 = 2435$

Step 7 - Square each treatment's sum of X (see step 2 above), divide each squared sum by its respective treatment sample population, and sum the three results.

$$\frac{(140)^2}{14} + \frac{(236)^2}{12} + \frac{(122)^2}{10} = 1400.00 + 4641.33 + 1488.40 = 7529.73$$

Step 8 - Subtract the result of step 5 from the sum obtained in step 7.

$7529.73 - 6889 = 640.73$

Step 9 - Subtract the result of step 8 from the result of step 6 .

$2435 - 640.73 = 1794.27$

Thus far, the following sum of scores for X has been generated:

Sum of Squares: X				
Between	7529.73	-	6889.00	= 640.73
Within	9324.00	-	7529.73	= 1794.27
Total	9324.00	-	6889.00	= 2435.00

[17] This procedure is a modified version of the analysis outlined in Bruning, James L. and Kintz, B.L., Computational Handbook of Statistics, Glenview, Ill: Scott, Foresman, and Co., 1968, pp. 173-177.

The same process will be followed for Y.

Step 10 - Square every Y score individually by treatment and sum across all three treatments (ΣY_g^2).

2560 + 7520 + 3074 = 13154

Step 11 - Add the sums of the Y columns (grand sum = Ty).

180 + 280 + 162 = 622

Step 12 - Square the grand sum of Ys and divide by the total population (N).

$$\frac{Ty^2}{N} = \frac{(622)^2}{36} = 10746.78$$

Step 13 - Subtract the result obtained in step 12 from the sum obtained in step 10.

13154 - 10746.78 = 2407.22

Step 14 - Square each treatment's sum of Y (see step 2), divide each squared sum by its respective sample population, and sum the three results.

2314.29 + 6533.33 + 2624.40 = 11472.02

$$\frac{(180)^2}{14} + \frac{(280)^2}{12} + \frac{(162)^2}{10} = 11472.02$$

Step 15 - Subtract the result of step 12 from the sum obtained in step 14.

11472.02 - 10746.78 = 725.24

Step 16 - Subtract the result of step 15 from the result of step 13.

2407.22 - 725.24 = 1681.98

Step 17 - Multiply each pretest score (X) by the corresponding posttest score (Y) in each treatment. Sum all the products across all treatments.

2056 + 6524 + 2426 = 11,006

Step 18 - Multiply the grand sum of X, obtained in step 4 by the grand sum of Y, obtained in step 11. Divide this product by N.

$$\frac{(498)\ (622)}{36} = 8604.33$$

Step 19 - Subtract the result obtained in step 18 from the sum obtained in step 17.

$$11006 - 8604.3 = 2401.7$$

Step 20 - Multiply each treatment's sum of X (see step 2) by that treatment's sum of Y and divide the product by that treatment's sample population and sum the three results.

$$\frac{(140)\ (180)}{14} + \frac{(236)\ (280)}{12} + \frac{(122)\ (162)}{10} =$$

$$1800 + 5506.7 + 1976.4 = 9283.1$$

Step 21 - Subtract the result of step 18 from the sum obtained in step 20.

$$9283.1 - 8604.33 = 678.77$$

Step 22 - Subtract the result obtained in step 21 from the result obtained in step 19.

$$2401.7 - 678.77 = 1722.93$$

Step 23 - Square the result of step 19, and divide by the result of step 6.

$$\frac{(2401.7)^2}{2435} = 2368.85$$

Step 24 - Subtract the result of step 23 from the result of step 13.

$$2407.22 - 2368.85 = 38.37$$

Step 25 - Square the result of step 22, and divide by the result of step 9.

$$\frac{(1722.93)^2}{1794.27} = 1654.43$$

Step 26 - Subtract the result obtained in step 25 from the result obtained in step 16.

$$1681.98 - 1654.43 = 27.55$$

Step 27 - The adjusted within group number of degrees of freedom =

N (Total population) - k (number of treatments) - 1 =

$$26 - 3 - 1 = 32.$$

Divide the result of step 26 by the adjusted within group degrees of freedom.

$$\frac{27.55}{32} = .86$$

Step 28 - Subtract the result of step 26 from the result of step 24.

38.37 - 27.55 = 10.82

Step 29 - The adjusted between group number of degrees of freedom =

k (treatments) - 1 = 3 - 1 = 2.

Divide the result of step 28 by the adjusted between group degrees of freedom.

$$\frac{10.82}{2} = 5.41$$

Step 30 - The F ratio = $\frac{MS_b}{MS_w} = \frac{S_b^2}{S_w^2}$, where

S_b^2 is the variance between groups and S_w^2 is the variance within the groups. Divide the result of step 29 by the result of step 27.

$$F = \frac{S_b^2}{S_w^2} = \frac{5.41}{.86} = 6.32 \qquad 6.32 > 5.34 \quad F \leq .05$$

$$df = \frac{k - 1}{N - k - 1} = \frac{2}{32}$$

Check for significance in F table. If significant proceed.

Step 31 - Adjust the treatment posttest means by the following formula:

$$\bar{Y}_k^{11} = b_w (\bar{X} - \bar{X}_k) + \bar{Y}_k$$

k = treatment

b_w = $\frac{\text{sum of products-within}}{\text{sum of squares: X-within}} = \frac{1722.93}{1794.27} = b_w = .960$

Total covariate mean ($\bar{X}$) can be obtained by dividing the sum obtained in step 4 by the total population.

$$\bar{X} = \frac{498}{36} = 13.83$$

The adjusted posttest means are found as follows:

$$\bar{Y}_1^{11} = .96\ (13.83 - 10) + 12.86 = 16.54$$

$$\bar{Y}_2^{11} = .96\ (13.83 - 19.67) + 23.33 = 17.73$$

$$\bar{Y}_3^{11} = .96\ (13.83 - 12.2) + 16.20 = 17.77$$

Apply the Scheffé Method to determine treatment effectiveness. (See ANOVA section)

Compare Y_1^{11} : Y_2^{11}

Compare Y_1^{11} : Y_3^{11}

Compare Y_2^{11} : Y_3^{11}

Summary for Analysis of Covariance (ANCOVA)

To apply the F ratio to the change in treatment means, required that the analysis of covariance be used to adjust the achievement means so that comparable differences could be analyzed. The obtained F equaled 6.32 and was significant, so the evaluator could infer that the type of treatment does influence the amount of achievement. The evaluator would proceed with the Scheffé method (described in the ANOVA section above) to isolate optimal treatments.

The Median Test for Two Correlated Samples

The median test (a sign test) is especially helpful when the same (correlated) target group is being administered a locally developed pretest and posttest. The observations (individual scores) are paired for each individual. Theoretically, there should be as many pupils increasing their scores (+) as there are decreasing their scores (-) if the pupils are the same at the end of the treatment as they were at the beginning. The difference between the total group's number of plus (+) signs and minus (-) signs leads to a z score. The z score is defined as follows:

$$z = \frac{|D| - 1}{N}$$

where $|D|$ is the absolute value of the total number of plus signs (+) minus the total number of minus signs (-) and

N = the number of sets of paired observations showing a sign change. (Do not use this statistical test unless at least 10 paired observations are available).

Illustration: An ESEA Title I remedial reading teacher had 22 primary pupils referred to her because of the pupils' diagnosed low word recognition abilities. The remedial reading teacher prescribed a treatment of vocabulary work, word analysis skills, etc. A paraprofessional tested (pre) the pupils in September with a word list composed of the 95 most frequently used nouns taken from the Dolch Basic 220 Word List. The same paraprofessional again tested (post) the 22 pupils in May with the same list. One point was awarded for each word correctly identified. Beneath each score in the table below is the sign of the change from the pretest to the posttest.

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Posttest	35	40	53	48	30	42	35	40	45	49	40	60	32	60	62	65	61	39	32	70	50	58
Pretest	20	21	35	54	18	48	35	30	43	39	40	38	27	67	60	70	51	20	26	32	50	62
sign of pretest-posttest	+	+	+	-	+	-	0	+	+	+	0	+	+	-	+	-	+	+	+	+	0	-

All zero changes are discarded. The sample group was reduced (by each discarded set) to N = 19. There were 14 plus signs and 5 minus signs.

$$z = \frac{|D| - 1}{\sqrt{N}}$$

$$D = 14 - 5 = 9$$

$$z = \frac{|9| - 1}{\sqrt{19}} = \frac{8}{4.36} = 1.84$$

Recalling that a z of 1.96 is required for significance at the .05 level and a 2.58 for significance at the .01 level (for nondirectional tests), the obtained z of 1.84 is not large enough to support the idea that the sample of pupils are significantly different in their knowledge of the 95 nouns at the end of the treatment. The sign test is a weak statistical test since it does not account for the magnitude of the change, but only the direction. (However, the use of this statistical analysis is a marked improvement over the earlier days in evaluation when the results would have been displayed as a 65 percent (14/22) improvement, when there actually is no significant improvement at all!)

The Median Test for Two Independent Samples

The median test belongs to the nonparametric family of sign tests. The median test is similar to the $\overline{z}$ ratio but does not require the same assumptions about uniform characteristics within a total "normal" population. When working with remedial learners, some evaluators choose the weaker sign tests since assumptions required for statistical tests for remedial learner populations are sometimes difficult to satisfy.

The median test for two independent samples rests on the premise that given a joint median for two samples selected from the same population, half of the population should score above the joint median and half below it. (A joint median is the median of the two samples considered as one group.) A 2 X 2 table is used to compute a value of chi-square. The Chi Square value of significance is used for inferences. (The chi-square statistic is discussed on the next page and again on page 88.)

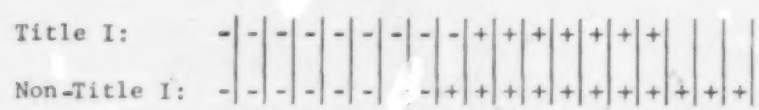
Illustration: A school contained classes of learners who were evincing characteristics of being dropout-prone. In the seventh grade, two classes of dropout-prone pupils were eligible for Title I funded activities, but only one class was selected to receive Title I treatments due to the size of the district's allocation. The Title I class teacher and the eligible but nonparticipating teacher both noticed that the pupils attitudes towards school (in general), instruction, teachers, and themselves (as learners) appeared highly negative. The Title I teacher constructed a behavioral checklist and administered it to the Title I class and the nonparticipating class in May of the school year. The rating was simply a score of 1 for every positive behavior and a 0 for every negative behavior.

The following are observations for the two independent classroom:

N (Title I) = 16 N (regular classroom) = 19

Title I Pupils:	9	9	9	11	12	14	14	15	15	18	19	20	20	21	22	23			
Non-Title I Pupils:	7	8	8	11	11	13	13	15	16	17	18	20	20	20	22	22	23	23	23

The median of N (Title I) + N (Reg. classroom) = 16. By assigning a plus (+) to each score at or above 16 and a minus (-) to each score below 16, the distribution looks like this:


Title I: - - - - - - - - - + + +

Several steps are required for this computation.

Step 1. Array the scores for each individual in pairs (i.e., pretest and posttest).

Step 2. Subtract one set of scores from the other in one direction (i.e., subtract the pretest from the posttest), to obtain different scores. If two scores are identical, delete this pair from the sets of scores.

Step 3. Rank the differences from the smallest to the largest by magnitude only (ignore the algebraic sign). If two differences are identical (tied for rank), assign an average rank to both scores as if the scores had been different (See Example.)

Step 4. Inspect the algebraic sign associated with each difference score to the left of the newly obtained rank. If it is positive, the rank is also positive.

Step 5. Sum the positive ranks and negative ranks separately.

Step 6. Assign the letter T to the smaller of the two sums (if no difference between samples exists, the two sums should be fairly close in magnitude while opposite in direction).

Step 7. Consult the tables for the "critical values of T in the Wilcoxon matched-pairs signed-ranks test" for the critical value of the N being observed. N = number of pairs of scores (which may have been reduced due to some paired observations of no difference). The table is contained in the appendix of George A. Ferguson, Statistical Analysis in Psychology and Education, New York: McGraw-Hill Book Co., 1966, p. 416.

Illustration A: A speech teacher had a rating scale of speech behaviors that were considered normal for entering first grade pupils. Pupils with dysfunctional speech patterns not attributable to physiological handicaps were selected for speech therapy. The teacher rated each pupil at the beginning of treatment and at the end of the treatment. (An alternate approach would have been to have taped pupil speech behaviors pretreatment and posttreatment and had the regional BOCES speech expert rate the behaviors).

The following are paired observations from pretestings and posttestings for 10 speech therapy pupils:

Pupil	1	2	3	4	5	6	7	8	9	10
Posttest	22	22	30	11	10	20	35	14	22	13
Pretest	11	15	33	11	5	22	20	10	8	20
d	+11	+7	-3	0	+5	-2	-15	+4	+14	-7
Rank	+7	+5.5	-2	0	+4	-1	-9	+3	+8	-5.5

Note: Pupil 4 exhibited no change, and is deleted. Hence, N=9 instead of 10.

Note: Pupils 2 and 10 are tied in magnitude (although not direction) occupying ranks 5 and 6. Hence, the tied rank of 5.5

The sum of the positive ranks is +27.5

The sum of the negative ranks is -17.5 ($T = |-17.5| = 17.5$).

The critical value of T when N=9 is equal to less than 6 for significance at the .05 level (two-tailed test) (taken from the Wilcoxon tables).

The target population is not significantly different on the posttest than it was on the pretest. The treatment can not be inferred to be effective for this target population. (In earlier years of Title I, evaluators would have (1) stated that the majority of pupils gained, and inferred that the treatment was effective, or (2) stated that the average gain was nearly two points, and inferred that the treatment was effective. As shown by the Wilcoxon matched-pairs signed-rank test, in actuality, no defensible inference acclaiming the effectiveness of the treatment can be made.)

For larger samples, a normal deviate z can be computed where

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

. The z must exceed the usual critical values of 1.96 for the .05 level, and 2.58 for the .01 for significance in a two-tailed test.

Chi-Square (X^2)

The chi-square statistic is used to compare whether the observed frequency of an event is the same as a theoretical or expected frequency. For remedial learners, the notion behind using the chi-square test is that if the observed frequency of a behavior differs from the expected frequency, then the particular sample of remedial learners under consideration no longer belongs to the population holding the expected frequency. Educators could infer that the intervention of the compensatory aid funded treatment contributed to the change in behavior for the sample. Illustration: For each of the past 5 years, the eighth grade class at the Malcolm X Junior High School had demonstrated a high rate of truancy. The average rate for district eighth grade remedial learners was approximately 10 days per year. During the 1969-70 school year, 71 percent

of the eighth grade class was truant more than the average. However, during the 1970-71 school year, only 47 percent of the pupils included in an Urban Education treatment were truant more than 10 days. The district evaluator wanted to know whether the difference between years was a minor fluctuation or whether it represented a significant difference. The evaluator chose to do a chi-square test.

Consider last year's class v. this year's class in absenteeism.

$N_1 = 70$ (last year) $\quad$ $N_2 = 85$ (this year)

The school average absenteeism rate is 10 days per year/pupil.

	absent) average	absent (average	
Last year's class (no treatment)	50 A	20 B	A & B = 70 (N_1)
This year's class (received Urban Ed. treatment)	40 C	45 D	C & D = 85 (N_2)
	A & C = 90	B & D = 65	

$$x^2 = \frac{N_{(\text{total 1 \& 2})} \quad (A \cdot D - B \cdot C)^2}{(A + B) \cdot (C + D) \cdot (A + C) \cdot (B + D)}$$

$$= \frac{155\ (2250 - 800)^2}{(70) \cdot (85) \cdot (90) \cdot (65)} = \frac{325887500}{34807500}$$

$$x^2 = 9.3+$$

df = (#rows - 1) (#columns - 1) = (2-1) (2-1) = 1

Critical value of chi-square for df=1 is 6.64 and can be obtained from the chi-square table found in the appendix of most books listed in the bibliography.

Since $9.3 > 6.64$, the last year's class is significantly different in terms of absenteeism than this year's class. (Also, $z = \sqrt{\chi^2}$, so $z = \sqrt{9.3} = 3.06$ which means that the two groups are over three standard deviations apart!)

Therefore, since the demographic data and beginning scores were assumed to be equivalent between the two years of classes, the evaluator could infer that the Urban Education treatment was having an effect on the truancy rate of the 1970-71 eighth grade class.

Describing Relationships Through Statistical Correlations

Statistical correlations are used to describe the degree of relationship between two or more known variables. In compensatory aid programs it is frequently desirable to find out what the relationship is between certain aspects of learning (and teaching) and achievement scores. If certain influences always appear when high achievement occurs, then evaluators can make certain inferences about what should be associated with a treatment to maximize learning. Although evaluators can not state flatly that certain influences cause high achievement, evaluators can recommend that teachers replicate certain events or conditions that have consistently appeared when remedial learners stopped falling burther behind their more advantaged peers. An illustration using the point biserial correlation is presented below to illustrate how to obtain or test for such a relationship.

Another use of the correlation approach is to try to establish a relationship between two variables (quantities that can assume different values) that remedial learners may possess or demonstrate. If one variable (i.e., achievement) coexists when another variable (i.e., aptitude) is present, then a relationship may exist. Such relationships (when known) permit evaluators to predict one variable in a learner from knowledge of the other variable. The Pearson Product-Moment correlation coefficient is used below to compute the theoretical value of such a relationship.

The omega-squared approach, also described below, is presented to illustrate how to test for the magnitude of relationship from a t score.

Pearson Product-Moment Correlation Coefficient

A correlation coefficient represents the relationship between two sets of scores. If each set of scores is considered to be a "variable," a classroom teacher might wish to compare a class on two "variables" to see if there is any correspondence.

The Pearson Product-Moment correlation coefficient is defined as

$$r = \frac{\Sigma xy}{N\,(SD_X)\,(SD_Y)}$$

where r = correlation coefficient

xy = the sum of the products of deviations above or below the group means $(\bar{X},\bar{Y})$

N = population size

SD = standard deviation of X

SD = standard deviation of Y

EXAMPLE: A remedial reading teacher felt that her Urban Education class of remedial learners (1) had a low generalized self-esteem due to repeated years of failure at academic activities, and (2) had far below grade level reading scores on the Metropolitan Achievement Test. During the year of individualized remedial reading instruction with the target class the teacher made a concerted effort (in cooperation with the other teachers) to foster a higher self-esteem in the class. The teacher had not only conducted the usual pretest/posttest administration for reading achievement, but had also administered (pre and post) a locally developed self-esteem measurement device. The teacher had applied t tests to both the differences between the anticipated posttest and obtained posttest means in reading, and the pretest and posttest means in self-esteem, and found "growth" over the period of instruction ($p \leq .01$ in both comparisons). However, the teacher wanted to know if there was any relationship between the amount of gain in reading and the amount of gain in self-esteem (i.e., Did the pupils who showed gains in reading also show gains in self-esteem?). The teachers arranged the data in the following table to compute the relationship between the reading scores and self-esteem scores.

N	X	Y	X^2	Y^2	XY
Pupil No.	Reading Score Gain (Mo.)	Self-Esteem Score Gain (pts.)			
1	10	20	100	400	200
2	8	17	64	289	136
3	9	19	81	361	171
4	8	16	64	256	128
5	11	22	121	484	242
6	10	19	100	361	190
7	9	17	81	289	153
8	7	13	49	169	91
9	7	15	49	225	105
10	8	15	64	225	120
11	9	21	81	441	189
12	10	20	100	400	200
13	11	22	121	484	242
14	12	24	144	576	288
15	6	10	36	100	60
16	8	16	64	256	128
17	9	20	81	400	180
18	10	21	100	441	210
19	11	22	121	484	242
20	10	19	100	361	190
N = 20	$\Sigma X = 183$	$\Sigma Y = 368$	$\Sigma X^2 = 1721$	$\Sigma X^2 = 7002$	$\Sigma XY = 3465$

The teacher then applied the following form of the Pearson Product-Moment correlation coefficient:

$$r = \frac{N \Sigma X \cdot Y - \Sigma X \Sigma Y}{\sqrt{\left[N\Sigma X^2 - (\Sigma X)^2\right]\left[N\Sigma Y^2 - (\Sigma Y)^2\right]}}$$, where X and Y are raw scores.

$$r = \frac{20\ (3465) - (183)\ (368)}{\sqrt{\left[20\ (1721) - (183)^2\right]\left[20\ (7002) - (368)^2\right]}}$$

$$r = \frac{69300 - 67344}{\sqrt{\left[34420 - 33489\right]\left[140040 - 135424\right]}}$$

$$r = \frac{1956}{\sqrt{4297496}} = \frac{1956}{2073.0} = +.94$$

The teacher interpreted the +.94 correlation coefficient as a high correlation (strong relationship) between gain in reading achievement scores and gain in self-esteem scores for this target group. (Note: the teacher cannot ascribe any part of the gain in reading to improved self-esteem or vice versa.) The teacher has evidence that the two scores for the same individuals vary in the same way. The teacher could graphically reconstruct this finding by making a correlation chart. The chart would visually confirm that the two "change" scores are related, and that variations in one score tend to go with variations in the other score. (cf. Guilford, pp. 91-112).

Use of the Point Biserial Correlation

The point biserial correlation (a product-moment correlation) is used when one variable is continuous (e.g., achievement scores) and the other variable is dichotomous (e.g., males and females). The point biserial correlation is often used by researchers who want to discriminate between two groups of pupils who received a treatment, but one group received an "extra something" that the other group did not receive.

The formula for the point biserial coefficient (r) is

$$r_{pt} = \frac{\bar{X}_p - \bar{X}_q}{SD_t} \cdot \sqrt{pq}$$

where SD_t = standard deviation of all the scores on the continuous variable. $SD_t = \sqrt{\frac{(X-\bar{X}_t)^2}{N}}$

p = proportion of individuals who received the "extra something" or is defined as discrete in one direction (e.g., male).

q = proportion of individuals who did not receive the "extra something" or is defined as discrete in another direction (e.g., female). (Note: p could equal the proportion of people who passed; q for those who failed a test.)

$\bar{X}_t$ = mean for the total group

$\bar{X}_p$ = mean for the "p" individuals

$\bar{X}_q$ = mean for the "q" individuals

Example: A social studies teacher taught a unit on Afro-American History to 28 black Title I participants in the Martin Luther King Junior High School. (The district had established a need to enhance the self-esteem of these target pupils). Due to unforeseen circumstances, only 12 of the target youngsters were able to attend a weekend field trip to an Afro-American History Museum that all of the target pupils were supposed to attend. At the conclusion of the unit, the teacher desired to know whether there was any correlation between those youngsters who attended the field trip and the youngsters scores on a locally developed self-esteem instrument. (The teacher had already applied a t test between pretest and posttest means for the group and found a significant increase). The self-esteem scores are continuous, while the participation in the field trip is dichotomous. The scores for the individuals are listed in the table below. Beneath the table the point biserial correlation is demonstrated.

TABLE FOR POINT BISERIAL CORRELATION

N INDIVIDUAL	X ESTEEM SCORE (posttest)	$X - \bar{X}_t$	$(X - \bar{X}_t)^2$	Attended field trip
1	10	-20	400	no
2	12	-18	324	yes
3	12	-18	324	no
4	14	-16	256	no
5	14	-16	256	yes
6	18	-12	144	yes
7	18	-12	144	no
8	20	-10	100	no
9	22	- 8	64	no
10	24	- 6	36	no
11	26	- 4	16	yes
12	26	- 4	16	yes
13	26	- 4	16	yes
14	28	- 2	4	no
15	30	0	0	no
16	31	+ 1	1	no
17	32	+ 2	4	yes
18	33	+ 3	9	no
19	35	+ 5	25	no
20	37	+ 7	49	no
21	40	+10	100	no
22	42	+12	144	yes
23	44	+14	196	yes
24	45	+15	225	no
25	46	+16	256	yes
26	48	+18	324	no
27	49	+19	361	yes
28	58	+28	784	yes
	$\Sigma X = 840$			

$N = 28$, p (yes) $= \frac{12}{28} = .43$ $\quad \Sigma \bar{X}_t = 30$ $\quad \Sigma(\bar{X}_t - X)^2 = 4578$

q (no) $= \frac{16}{28} = .57$

$$SD_t = \sqrt{\Sigma(X - X_t)^2 / N} = \sqrt{4578/28} = \sqrt{163.5} = 12.79$$

Mean score of field trip pupils = $\bar{X}_p = 32.75$

Mean score of nonfield trip pupils = $\bar{X}_q = 27.94$

$$pbi = \frac{\bar{X}_p - \bar{X}_q}{S_t} \sqrt{pq} = \frac{32.75 - 27.94}{12.79} = \sqrt{(.43)(.57)} = .19$$

At least one inference possible from this coefficient of .19 is that the field trip experiences (and expense) can not be defended on the basis of improving self-esteem as measured by this instrument. However, before stating such a conclusion, the usual practice is to test the obtained r_{pbi} (.19 in this case) for significance by means of the t ratio test. The evaluator is testing the hypothesis that r_{pbi} is really zero, and that the value of .19 is simply a number obtained by chance (sampling error).

$$t = r_{pbi}\sqrt{\frac{N-2}{1-r^2_{pbi}}} \quad \text{where the df} = N-2.$$

$$t = .19\sqrt{\frac{28-2}{1-(.19)^2}} = (.19)\cdot(5.19) = .99. \quad df = 26$$

Since the critical value of t for $p \leq .05$, df = 26, is 2.056 the computed t of .99 was not significant. The computed r_{pbi} of .19 is not different from zero and is indicative of no relationship between these field trip experiences and self-esteem as measured by the locally developed instrument.

Use of the t Statistic to Account for the Treatment Impact

Title I evaluators sometimes use an obtained t ratio from uncorrelated samples to estimate the amount of effect that a treatment actually made on the target group. In other words, evaluators desire to know not only whether a treatment contributes to a significant increase in achievement, but also how much of the increased achievement can be associated with the treatment. By paying more attention to the association between treatment effectiveness in addition to the outright significantly increased achievement, decision makers can select from among the better treatments for future implementation with remedial learners.

William Hays[18] suggests a way of estimating the strength of the association between the difference between means for two uncorrelated samples and the treatment effect. Hays defines the association as omega squared (ω^2) which is translated as the percent that the treatment accounts for in the variance of the obtained score.

$$\text{est } \omega^2 = \frac{t^2 - 1}{t^2+N_1+N_2 - 1}$$

where est ω^2 = strength of the association
t = the usual t ratio
N_1, N_2 = the uncorrelated sample sizes

---

[18] Hays, William L. Statistics for Psychologists, New York: Holt, Rinehart, and Winston, 1963, pp. 327-328.

Illustration: Fifty-six fourth grade remedial learners in reading who had scored below the 23rd percentile on the New York State PEP Test were randomly assigned to the regular classroom or the Title I reading treatment class.

The mean, $(\bar{X}_1)$ of the treatment group ($n_t = 28$) was 35.

The variance was 7, $SD_{\bar{X}_1} = 2.6$, $SE_{\bar{X}_1} = .50$.

The mean $(\bar{X})_2$ of the regular classroom ($n_t = 28$) was 33.

The variance was 5, $SD_{\bar{X}_2} = 2.2$, $SE_{\bar{X}_2} = .42$.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{(SE_{\bar{X}_1})^2 + (SE_{\bar{X}_2})^2} = \frac{35-33}{(.5)^2 + (.42)^2} = \frac{2}{.65} = 3.07$$

The critical value of t needed at $p \leq .05$ was 2.68. Since the computed 3.07 was greater than 2.68, the difference is significant ($p \leq .05$). An association probably does exist.

Hay's estimate: $\omega^2 = \frac{(3.07)^2 - 1}{(3.07)^2 + 28 + 28-1} = .13$

Using Hays' estimate then, the treatment appears to account for only about 13 percent of the obtained score.

Illustration: Consider the same population characteristics as just given, except reduce the two samples to 10 and reduce the variances by three-fourths.

Title I treatment	Regular Classroom
$\bar{X}_1 = 35$	$\bar{X}_2 = 33$
$S_1^2 = 1.75$	$S_1^2 = 1.25$
$SD_1 = 1.32$	$SD_1 = 1.12$
$SE_{\bar{X}_1} = .44$	$SE_{\bar{X}_2} = .37$
$N_1 = 10$	$N_2 = 10$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{(SE_{\bar{X}_1})^2 + (SE_{\bar{X}_2})^2} = 3.45$$

Again the t ratio is significant at $p \leq .01$. An association probably does exist.

Hay's estimate: $\omega^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1} = \frac{(3.45)^2 - 1}{(3.45)^2 + 10 + 10 - 1} = .352$

In this case, the treatment appears to account for over 35 percent of the variance in the score obtained.

The strength of the association is sharply contrasted by the two examples given above. Both illustrations were significant at $p \leq .05$. The difference between the means was the same. An association between treatment type (independent variable) and variance in the score obtained (dependent variable) existed in both cases. However, the second case demonstrated a much stronger association than existed in the first case. A project coordinator can use such inferences for his selection of remedial reading treatments with greater assurance of the effects of the treatment itself.

Comments to Coordinators

Many of the statistical techniques employed in this handbook will appear to coordinators to be unnecessarily rigorous when applied to data collected from categorically aided learners. Supplements to this handbook may eventually contain simpler procedures to estimate behavioral change. However, at this point in time, the techniques contained in this handbook have been approved by the experts in the field of evaluation.

For coordinators who might like to review the underlying assumptions connected with distributions, the robustness of certain statistical tests, concepts associated with variance, etc., a bibliography is provided. Most of the references contain the tables of the critical values associated with several statistical techniques. Particularly recommended for novice coordinators are Guilford's "Fundamental Statistics in Psychology and Education" for parametric techniques, and Siegel's "Nonparametric Statistics for the Behavioral Sciences."

Appendices A, B, and C are samples of ESEA Title I component evaluations approved during fiscal 1970-71. Admittedly, the statistics tend to be simple and mostly of the t ratio variety. However, many upstate districts are just at the verge of assembling data to establish behavioral change (i.e., rate of achievement) prior to treatments funded by categorically aided programs and during such treatments. This handbook and its future supplements should assist in that major step of preliminary data analysis.

Appendix D is a flow chart for coordinators to use in planning for project evaluation. In a sense, the flow chart serves as a planning model to be followed when constructing that section in a project proposal that calls for evaluation methods. A coordinator who follows the model and submits adequate responses to each step in the chart sequence would be unlikely to have his project rejected by the funding source for insufficient information concerning proposed evaluation methods. The Bureau of School and Categorical Programs Evaluation is the appropriate source of information if a coordinator would like clarification of the evaluation planning model, or other technical assistance.

APPENDIX A - Instructional Activity

ESEA Title I project proposal #58-02-11-72-001, Middle Country, N = 840, $233,215.20.
Contact Person: Dan Birecree

The following evaluation design was abstracted from a Middle Country project application, sections three and four.

The thrust of the component was to increase the individualization of reading instruction for categorically aided learners. Aides were added in four categories of functions including primarily direct instruction or support services (thereby freeing the classroom teacher for more direct instruction). The district desired to know how to use aides most effectively to increase pupil achievement.

Pupil Performance

Objectives: The target population will demonstrate a significant increase beyond expectation in reading achievement as measured by the Stanford Achievement Test.

Sampling Procedure: A plan of cluster sampling in all schools will be followed by a within-school random stratified selection by grade level (to exceed 120 pupils).

Design: A rate of growth design will be employed. The pupils will be pretested and posttested with the Stanford Achievement Test. An anticipated growth rate will be contrasted with the actual growth rate. In addition, the predicted posttest mean (predicted from the pretest) will be compared to the posttest mean obtained from the second administration of the Stanford Achievement Test.

Measuring Devices: Stanford Achievement Test administered in October and May.

Data Analysis: t ratio test of correlated means will be applied to the (1) rate of growth scores and the (2) predicted posttest and actual posttest. Analysis of variance will be applied using the rate of growth as the dependent variable and the following four treatment (independent) variables.

a. Aides used largely for direct instruction in one classroom

b. Aides used largely for noninstructional support activities in one classroom

c. Aides used largely for direct instruction across several classrooms

d. Aides used largely for noninstructional activities across several classrooms

Aide Performance

Objectives: The aides will demonstrate a level of adequate or better performance as measured by the attached locally developed checklists.

Sampling Procedure: All aides

Design: A fall, midwinter, and spring observation will be conducted by the coordinator and/or supervisor in each of the four aide categories.

Measuring devices:
1. observor rating checklist
2. teacher rating schedule (aide performance)
3. aide self-evaluation schedule

Data Analysis:
1. Correlation (Pearson) between observor ratings and teacher ratings.
2. Aide ratings by observors will be tested for significant differences (t ratio) if an examination indicates changes (in either direction) between fall and spring observations (by category).

Possible category: An analysis of items rated low by teachers, aides, or observors will generate specific areas for in-service training or preservice orientation for aides and/or teachers.

APPENDIX B - Support Services

ESEA Title I project proposal #66-23-00-72-001, Yonkers, N = 3730, $1,161,450
Contact Person: Joan Chertok

The following evaluation design was abstracted from the Yonkers project application sections three and four. Only the support service component evaluation design of project "ACTION" is presented below.

The district decided to expend part of its allocation on guidance counselors, social workers, psychologists, and attendance officers. The district had determined that categorically aided learners (1) had a high rate of absenteeism, (2) had a noticeable hostile or apathetic attitude toward school, (3) had a low expectation for success in school, and (4) frequently "acted out" or were disruptive in class when they did attend. The district desired to know whether the supplementary services cutting across several grades changed nonacademic behavior.

Objective 1: The target population will demonstrate a significant reduction in truancy.

Sample: A total sample greater than 120 pupils will be randomly selected from target classrooms within each school.

Design: A pretreatment mean attendance rate will be established for the sample, based on the pupils' last year's attendance record as taken from the permanent record cards. A treatment mean attendance rate will be obtained from this year's attendance patterns.

Data Analysis: A correlated t test or z ratio will be computed ($p \leq .05$).

Objective 2: The target population will demonstrate a significant improvement in attitude toward school as measured by a locally developed school attitude instrument (attached).

Sample: Stratified, clustered random sample. $N < 120$. A nontreatment group of eligible nonparticipants will be selected under the same procedure.

Design: A preattitudinal and postattitudinal survey will be administered in September and May.

Data Analysis: An analysis of covariance will be employed.

Objective 3: The target population will demonstrate an increased expectancy for success in school as measured by the locally developed instrument.

Sample: Stratified by grade, clustered by school, randomly selected. $N < 120$.

Design: A pretest will be administered in September. A posttest will be administered in June. The incidence of participation at the parent involvement sessions will be recorded.

Data Analysis: A t ratio will be applied to the correlated means to determine whether a significant ($p \leq .05$) change in expectation has occurred. An analysis of covariance will also be applied to the pupils' pre and post scores. The two categories for treatment analysis are the pupils whose parents were concerned and attended the involvement sessions; and, the pupils whose parents may or may not have been concerned who did not attend the involvement sessions. The pupil scores on the school success expectancy instrument will serve as the dependent variable.

Objective 4: The target population will demonstrate a significant reduction in disruptive behavior as measured by referral for discipline records.

Sample: A random sample of 50 pupils from the pool of pupils selected for guidance, social work, and psychologist intervention.

Design: The mean rate of referral based on last year's records will be compared to the mean rate of referral for disciplinary action this year by means of correlated t ratio.

APPENDIX C - Instructional Activity (Summer)

ESEA Title I project proposal #44-18-00-71-002, Port Jervis, N = 130, $16,025,
Contact Person: H. Edward Dux

The following project proposal from Port Jervis was a summer project devoted to intensive remedial reading activities. It was the district's intent to average nearly 5 months' reading comprehension achievement in approximately 5 weeks of instruction for over 300 pupils. The objectives and evaluation plan (Section III) are reproduced below exactly as they were received.

A. Objectives: To raise the reading comprehension of one-third of the students by one grade level (+1.0); to raise the reading comprehension of one-third of the students by half a grade level (+.3); to raise the reading comprehension of one-third of the students by 1 month (+.1) all within the specified 5 week program. The criteria used for achieving the above stated objectives are:

1. Temple University's Individual Reading Inventory (see attached).

2. Standardized diagnostic reading test published by Education Progress Corporation; 8538 E. 41st Street, Tulsa, Oklahoma.

B. Development and Application of Sampling Plan: Target population -- all students attending the program, grades 1 through 8 utilizing randomized sampling.

C. Methodology and Management:

1. Pretesting and posttesting using Temple IRI in grades 1 through 8.

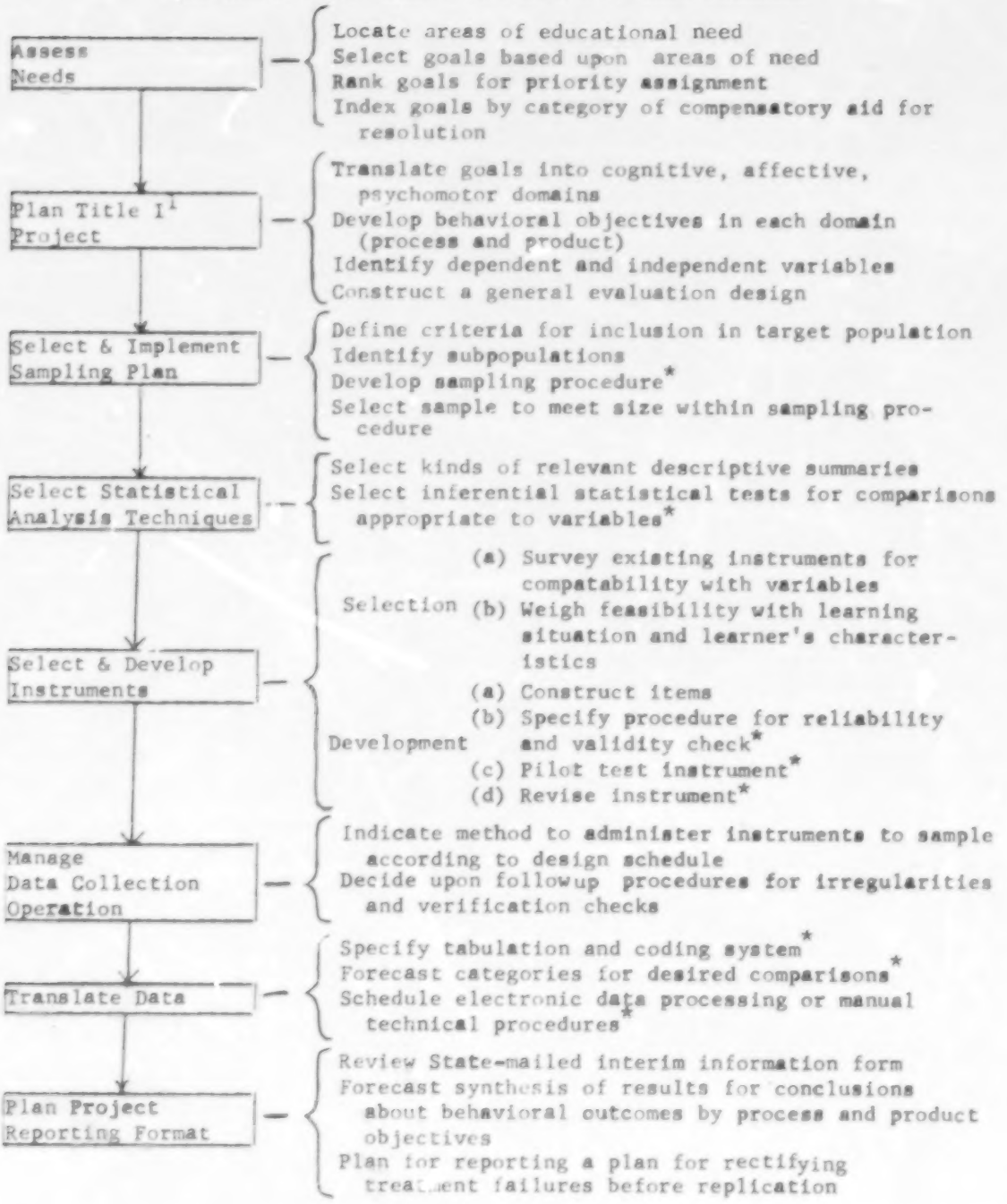
2. Pretest and posttest using standardized diagnostic reading test published by Educational Progress Corporation will be given to one randomized section in each grade 1 through 8.

3. As a means of judging the merits of the Education Progress Corporation materials, the test results from the randomized sections of grade 3 will be compared. All sections of grade 3 will receive pre- and post- Temple IRI's. In addition, one section of grade 3 will receive prediagnostic and postdiagnostic reading tests through E.P.C. This randomized section will utilize only the E.P.C. Reading Series and Language Arts series materials. The other third grade sections will use regular summer school materials as in the past. They will not use E.P.C. materials.

Data Analysis: Analysis of pre- and post- IRI together with the E.P.C. diagnostic reading pretests and posttests will be compared based on reading comprehension measured in terms of monthly growth.

Treatment by statistical techniques for inferential conclusions will be checked by means of the t test.

EVALUATION FLOW CHART FOR TITLE I PROJECT PLANNING


Assess Needs
Locate areas of educational need
Select goals based upon areas of need
Rank goals for priority assignment
Index goals by category of compensatory aid for resolution
Plan Title I[1] Project
Translate goals into cognitive, affective, psychomotor domains
Develop behavioral objectives in each domain (process and product)
Identify dependent and independent variables
Construct a general evaluation design
Select & Implement Sampling Plan
Define criteria for inclusion in target population
Identify subpopulations
Develop sampling procedure*
Select sample to meet size within sampling procedure
Select Statistical Analysis Techniques
Select kinds of relevant descriptive summaries
Select inferential statistical tests for comparisons appropriate to variables*
Select & Develop Instruments
Selection
(a) Survey existing instruments for compatability with variables
(b) Weigh feasibility with learning situation and learner's characteristics
Development
(a) Construct items
(b) Specify procedure for reliability and validity check*
(c) Pilot test instrument*
(d) Revise instrument*
Manage Data Collection Operation
Indicate method to administer instruments to sample according to design schedule
Decide upon followup procedures for irregularities and verification checks
Translate Data
Specify tabulation and coding system*
Forecast categories for desired comparisons*
Schedule electronic data processing or manual technical procedures*
Plan Project Reporting Format
Review State-mailed interim information form
Forecast synthesis of results for conclusions about behavioral outcomes by process and product objectives
Plan for reporting a plan for rectifying treatment failures before replication

[1]Activity (treatment) planning, omitted from this chart, occurs concurrently after the behavioral objectives are specified.

*Consultant/contractor help advisable.

Bibliography

Brunning, James L. and Kintz, B. L. Computational Handbook of Statistics. Glenville, Ill.: Scott, Foresman, and Co., 1968.

Ferguson, George A. Statistical Analysis in Psychology and Education. 2nd ed. New York: McGraw Hill Book Company, 1966.

Games, Paul A. and Klare, George R. Elementary Statistics: Data Analysis for the Behavioral Sciences. New York: McGraw-Hill Book Co., 1964.

Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs: Prentice-Hall, Inc., 1968.

Guilford J. P. Fundamental Statistics in Psychology and Education. 4th ed. New York: McGraw-Hill Book Co., 1965.

Hays, William L. Statistics for Psychologists. New York: Holt, Rinehart, and Winston, 1963.

Horst, Donald P., Tallmadge, G. Kasten, and Wood, Christin T. Measuring Achievement Gains in Educational Projects (RMC Report UR-243). Los Altos, California: RMC Research Corporation, 1974.

McGuigan, Frank J. Experimental Psychology: A Methodological Approach. 2nd ed. Englewood Cliffs: Prentice Hall, 1968.

Seigel, Sidney. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co., 1956.

Young, R. K. and Veldman, J. D. Introductory Statistics for the Behavioral Sciences. New York: Holt, Rinehart, and Winston, Inc., 1965.